Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Sistema de Previsions Horàries d'Energia Eòlica basat en Màquines de Vectors de Suport

MEMÒRIA

Autor: Gonzalo Espinosa Duelo
Director: Josep Anton Sànchez

Convocatòria: Juliol 2016

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

En aquest treball es descriu el sentit estratègic de realitzar previsions de la producció d'energia eòlica i es proposa un sistema de predicció a curt termini de la potència elèctrica horària generada per un parc eòlic basat en els models no paramètrics desenvolupats originalment per Vladimir Vapknik: les màquines de vectors de suport. Concretament es desenvolupen estratègies basades en l'aplicació d'aquests algorismes per a regressió. D'aquesta manera, s'avalua la precisió dels models utilitzant únicament variables predictores numèriques de les produccions del parc eòlic, sense tenir en compte les condicions meteorològiques o altres factors físics de l'entorn.

Sumari

1. Introducció

En aquesta secció es descriu primerament l'objecte d'estudi i el propòsit del treball. Posteriorment, s'exposa el pla de treball que s'ha seguit per desenvolupar el projecte.

1.1. Motivació i objectius

Actualment, a la península ibèrica, la predicció de producció energètica de parcs eòlics té un caràcter estratègic molt important en la situació del mercat elèctric, ja que el preu de l'energia eòlica depèn estrictament de la seva producció diària. Poder predir la seva generació permet gestionar el sistema elèctric de forma eficaç, proporcionant capacitat de previsió per tal de poder efectuar ofertes ajustades a la demanda. A més, la impossibilitat d'emmagatzemar l'energia a gran escala implica que conèixer les quantitats d'energia que es produiran sigui una qüestió d'encara major importància.

Per tal crear el model predictiu de producció d'energia eòlica a curt termini, es proposa implementar les tècniques d'aprenentatge automàtic supervisat, concretament màquines de vectors de suport. Aquest conjunt d'algorismes, van originalment dissenyats per a la solució de problemes no lineals de classificació però que, a causa de la seva capacitat de generalització, recentment s'ha aplicat a problemes de regressió i predicció de sèries temporals.

L'objectiu que motiva el desenvolupament d'aquest treball és construir un sistema de predicció a curt termini de la potència elèctrica horària que produeix un parc eòlic, a partir de dades de produccions horàries, basat en màquines de vector de suport.

L'abast del treball és explorar i avaluar l'eficàcia de les màquines de vectors de suport aplicades a la predicció de series temporals de producció d'energia eòlica. Es proposarà algunes estratègies per afrontar el problema de previsions, així com una metodologia per a la selecció de les variables que controlen les màquines de vectors de suport i es valorarà la precisió de les prediccions que ens aporta aquesta tècnica d'aprenentatge automàtic.

1.2. Pla de treball

Per a dur a terme aquest projecte, s'ha realitzat un pla de treball on es desglossen les diferents etapes del cronograma que s'ha seguit per a fer possible l'acompliment dels objectius, tenint en compte que es disposen de poc més de quatre mesos per a presentar l'informe finalitzat.

Proposta	Febrer
Entendre el conflicte	Febrer
Familiarització amb Machine Learning i R	Febrer
Comprendre els principis de SVM	Març
Processament de dades	Març
Model bàsic SVR: entrenament	Març
Validació del model: test	Abril
Millora del model	Abril
Anàlisi de resultats i conclusions	Maig
Estudi econòmic i medi ambiental	Maig
Ultimar el redactat	Juny

2. Producció d'Energia Eòlica

En aquesta secció es presenta l'energia eòlica i la seva producció, s'explica la seva importància al nostre país i es justifica el valor que té per al mercat elèctric espanyol fer prediccions a curt termini de produccions energètiques.

2.1 L'energia eòlica

L'energia eòlica genera electricitat a través de la força del vent, mitjançant la utilització de l'energia cinètica produïda per efecte dels corrents d'aire. Es tracta d'una font d'energia neta i inesgotable, sent l'energia renovable més madura i desenvolupada fins al moment.

Des de principis del segle XX, produeix energia a través dels aerogeneradors. L'energia eòlica mou una hèlix i, mitjançant un sistema mecànic, fa girar el rotor d'un generador que produeix energia elèctrica. Els aerogeneradors solen agrupar-se en concentracions denominades parcs eòlics per tal d'aconseguir un millor aprofitament de l'energia, reduint seu impacte ambiental.

L'eòlica constitueix una garantia de sostenibilitat mediambiental, ja que no contamina, és inesgotable, evita emissions CO_2 i frena l'esgotament de combustibles fòssils, contribuint a evitar el canvi climàtic.

A més, el sector eòlic és clau per complir els objectius europeus de consum d'energia a través de fonts renovables el 2020, fet que consolida la eòlica com a aposta estratègica i energia del futur.

2.2 La producció d'energia eòlica a Espanya

La indústria eòlica espanyola és un referent mundial. En els últims anys, la indústria d'energia eòlica ha tingut una evolució espectacular. L'elevat desenvolupament tecnològic i econòmic de les empreses espanyoles del sector els ha permès un reconeixement mundial important i una presència cada vegada més gran en els principals mercats internacionals.

D'altra banda, atretes per la notable activitat industrial nacional, empreses d'altres països s'han instal·lat a Espanya, el que contribueix al creixement econòmic del país i suposa un impuls més en l'avanç tecnològic, alhora que també destaca la presència de moltes de les nostres empreses en altres països. A Espanya, es disposa d'un conjunt empresarial rellevant en totes les fases de la cadena de valor del sector i amb una clara orientació cap a un mercat global (s'exporta tecnologia per valor de més de 2.000 milions d'euros a l'any).

Aquesta xarxa d'empreses està composta tant promotors de parcs eòlics i fabricants d'aerogeneradors, com per empreses de fabricació de components i serveis. Actualment, més de 20.000 persones treballen en el sector al nostre país.

Espanya, amb una capacitat eòlica instal·lada de 22.988 MW a tancament de l'any 2015, és el cinquè país del món, pel que fa a potència eòlica instal·lada, superat per la Xina, Estats Units, Alemanya i recentment, la India. La producció d'energia elèctrica a través d'aquest tipus d'instal·lacions cobrint en l'any 2015 un 19,4% del consum elèctric del país, sent la tercera tecnologia generadora d'electricitat 47.704 GWh generats.

2.3 El mercat d'energia elèctrica

La importància d'establir un model predictiu de la producció d'energia elèctrica provinent dels aerogeneradors recau, en primera instància, en el funcionament del mercat energètic. Aquest mercat es basa en la compra-venda majorista i minorista de totes les energies elèctriques, incloent l'energia eòlica, que no es poden emmagatzemar de manera eficient en grans quantitats i que estan restringides a certes particularitats que caracteritzen el sector.

El mercat elèctric a Espanya engloba tot el conjunt de mercats on es negocia la compra i venda d'energia elèctrica de la península. Es va establir com a conseqüència de la liberalització del sector elèctric que va tenir lloc en l'any 1997. Fins a l'any 1997 el sistema elèctric espanyol estava estructurat com un sistema regulat en el qual el Govern establia el preu de l'electricitat, que remunerava la totalitat dels costos incorreguts (generació, transport i distribució de l'electricitat) a un conjunt de companyies elèctriques privades. Actualment, es basa en el reconeixement de dos tipus d'activitats: activitats parcialment liberalitzades (generació i comercialització) i activitats regulades (transport i distribució).

El conjunt d'activitats pertanyents al mercat elèctric està constituït per dos sectors: el mercat majorista i el mercat minorista. El mercat majorista distingeix dues formes de contractació: Per una banda, mitjançant la competència directa entre els venedors i compradors a través del Mercado Ibérico de Electricidad (MIBEL). D'altra banda, existeix la possibilitat de contractació lliure entre els generadors i compradors qualificats, però que ha de ser comunicada a l'OMIE que actua com organisme regulador.

2.3.1 MIBEL: mercat diari i intradiari

El mercat diari, o del dia abans, és el mercat on es negocia la major part del volum energètic. Es realitza a les 12.00 del dia D-1, on els productors i compradors han d'haver presentat les seves ofertes (en quantitat d'energia i preu) per a cada franja horària del dia D.

El mercat intrahorari està caracteritzat per sis sessions durant el dia D on els productors i compradors poden realitzar noves ofertes per a la negociació durant les properes franges del dia. Així, els productors ajusten les seves previsions de produccions amb dades més properes i fiables i els consumidors ajusten les seves ofertes als nous programes de consum actualitzats.

Tant en el mercat horari com en el intrahorari, el MIBEL caça un preu a partir del preu mínim que al que estan disposats a vendre els productors i el preu màxim que estan disposats a pagar els consumidors, alhora que garanteix la quantitat total acceptada de compra és igual a la quantitat total acceptada de venda. Així, totes les vendes o compres d'energia es subhasten, per a cada hora, al preu de cassació obtingut per a aquesta hora.

Cal afegir, que els productors sovint també es converteixen en compradors al mercat, ja que quant no són capaços de produir tota la quantitat que han venut, l'han de recomprar per tal d'evitar grans penalitzacions.

2.3.2 La importància de la predicció al mercat energètic

El propi caràcter predictiu del mercat energètic és doncs el motiu pel qual la predicció de produccions energètiques sigui tan important. Des del punt de vista de generació eòlica, o de qualsevol altra font d'energia renovable, la seva previsió és útil tant per a l'operador del sistema com per als agents del mercat o els propietaris de parcs. L'operador del sistema elèctric necessita conèixer amb antelació suficient la quantitat d'energia eòlica que serà injectada a la xarxa per gestionar la potència que hauran de generar les centrals convencionals, amb l'objectiu de cobrir la demanda total del sistema. Mentrestant, els agents de mercat (compradors i venedors) estaran interessats a conèixer amb la major certesa possible la potència que generaran els seus parcs eòlics amb l'objectiu de seguir les estratègies que resultin més rendibles en el mercat d'energia elèctrica.

2.4 La predicció de producció d'energia eòlica

L'eòlica és una forma de generació no programable, ja que només es produeix energia quan bufa el vent, que pot arribar a ser molt variable fins i tot en el curt termini, amb possibilitat d'intermitència i grans canvis en intervals curts de temps. Això fa que sigui difícil conèixer amb antelació i precisió suficient la quantitat d'energia eòlica amb la qual es pot comptar en cada moment. Inevitablement les previsions són afectades per errors o incerteses de predicció.

Si el vent disminueix, la potència generada als parcs eòlics també disminueix, i aquesta falta de potència ha de ser reemplaçada per altres fonts de generació per a que la demanda

elèctrica no es vegi afectada. En altres ocasions, pot passar que no es pugui integrar en el sistema tota la producció eòlica disponible, ja que l'energia eòlica no es genera d'acord a les necessitats de consum, i sigui necessari reduir el subministrament d'aquesta font d'energia. Per tot això, la predicció de generació eòlica s'ha convertit en un tema clau per fer factible el desenvolupament i implantació de l'energia eòlica, i la seva integració en el sistema elèctric.

Donada la importància dels sistemes de prediccions en el mercat elèctric, al llarg dels anys s'han desenvolupat diverses tècniques i algorismes per a estimar produccions a curt termini. Concretament, podem distingir dues formes de modelització predictiva de produccions d'energia eòlica: els models físics i els models numèrics.

Els models físics tenen en compte consideracions físiques per adaptar les prediccions de vent a la zona concreta de l'emplaçament del parc. S'utilitzen models meteorològics per tal de determinar la velocitat del vent incident en les turbines del parc i posteriorment calculen la predicció de potència per mitjà de la corba de potència.

Els models numèrics consisteixen bàsicament en analitzar els valors passats d'una variable i d'altres variables relacionades per buscar patrons significatius, amb l'objectiu de poder conèixer o extrapolar els valors que prendrà aquesta variable en el futur. És on trobem els models paramètrics clàssics de series temporals. A més dels models estadístics clàssics paramètrics, existeixen un conjunt de models numèrics freqüentment utilitzats: les tècniques d'aprenentatge automàtic. En aquest treball és desenvoluparà una d'aquestes tècniques, les màquines de vectors de suport.

3. Màquines de vectors de suport

En aquesta secció, es descriuen els fonaments teòrics de les màquines de vectors de suport. Primerament s'introdueix la família a la què pertany: el Machine Learning. Posteriorment s'expliquen els principis de les màquines de vectors de suport i les seves dues aplicacions, amb l'objectiu de tenir una noció bàsica dels conceptes que després seran implementats i entendre aquest conjunt d'algorismes que típicament es pren com a caixa negra.

3.1 Què és machine learning?

El *machine learning* o aprenentatge automàtic és una branca de la ciència de la intel·ligència artificial que té per objectiu desenvolupar tècniques que permetin a les màquines aprendre. Concretament, es tracta de un conjunt de eines de complexitat computacional que permeten generalitzar comportaments a partir d'informació no estructurada.

Tot i que el machine learning és un concepte més aviat d'enginyeria, sovint es solapa amb el camp de l'estadística, ja que està estrictament relacionat amb l'anàlisi de dades. De fet, pretén descobrir patrons ocults, correlacions desconegudes, trobar informació útil de les dades i –fet pel qual ens interessa especialment– és útil per realitzar anàlisis predictiu. Un forma d'entendre que és l'aprenentatge automàtic és concebre'l com "fer estadística des de la informàtica".

3.2 Mètodes d'aprenentatge

Per poder dur a terme els processos intuïtius que persegueix, el *machine learning* engloba un conjunt de mètodes que es basen en l'aprenentatge a partir d'entrenaments (dades d'entrada) pels quals s'ajusten certs paràmetres globals i es caracteritzen els objectes (dades de sortida). Aquests mètodes d'aprenentatge comprenen dues branques d'algorismes: els d'aprenentatge supervisat i no supervisat.

Tot i que l'aprenentatge engloba tant classificació com regressió de dades, els conceptes de supervisat i no supervisat van néixer de la mà de la classificació i la diferència té molt a veure en la seva aplicació. Bàsicament, en els algorismes supervisats es proporciona l'objecte durant el seu entrenament, mentre que en els no supervisats l'objectiu a assolir no està present. És a dir, mentre que la classificació supervisada es té un coneixement a priori de les diferents classes, en la no supervisada no se'n té i les classes son proporcionades per els propis algorismes.

3.2.1 Aprenentatge no supervisat

L'aprenentatge no supervisat fa servir dades històriques que no estan etiquetades amb la finalitat d'explorar-les per trobar alguna estructura o forma d' organitzar-les. Engloba doncs un conjunt de eines de classificació, per a resoldre problemes on es té un conjunt de observacions X amb absència de Y (etiqueta o classe pertinent).

Actualment és molt freqüent el seu ús en algunes empreses per agrupar els seus clients amb característiques o comportaments similars, per tal de fer campanyes de màrqueting altament segmentades.

Alguns dels exemples de mètodes d'aprenentatge no supervisat més comuns són:

- Tècniques d'Agrupament (Clustering)
- Anàlisi discriminant lineal (LDA)
- Anàlisi discriminant quadràtic (QDA)
- Anàlisi discriminant logístic
- Xarxes neuronals (ANNs)
- Mapes autoorganitzats (SOMs)

3.2.2 Aprenentatge supervisat

L'aprenentatge supervisat se sol utilitzar per a la predicció basada en comportaments o característiques observats en conjunts de dades d'entrenament. A partir d'un conjunt de dades on es coneix (X,Y), l'algorisme d'aprenentatge supervisat extreu una funció de decisió per predir Y per a una nova observació X. Si la sortida de la funció de decisió és un valor discret o etiqueta de classe aquesta funció es coneix com a funció de classificació. Altrament, si la sortida és contínua la funció amb sortida numèrica es diu funció de regressió. És a dir, a diferència de l'aprenentatge no supervisat, proporciona eines que es poden utilitzar tant per a classificació com per a regressió, fet que interessa especialment per a la realització d'aquest treball.

Alguns dels exemples de mètodes d'aprenentatge supervisat més freqüents són:

- Xarxes neuronals (ANNs)
- Màquines de vectors de suport (SVM)
- Nearest-neighbor classiers
- Linear least squares t (LLSF)
- Classificador Bayesià Ingenu (NB)

3.3 Principis del SVM

3.3.1 Introducció a les Màquines de Vectors de Suport

Les màquines de vectors de suport, conegudes per la seva denominació en anglès Support Vector Machine (SVM), representa una poderosa tècnica d'aprenentatge que s'utilitza per a classificació, regressió i detecció de valors atípics. Va ser introduïda per Vladimir Vapnik [2] [3] als anys 90 i actualment s'ha consolidat en la seva aplicació per a establir models predictius. Tot i que van ser desenvolupades originalment per a classificació, en el nostre cas ens interessa especialment per la seva aplicació de regressió, que utilitzarem com a eina predictiva de produccions energètiques a un parc eòlic.

Concretament, SVM són un conjunt d'algorismes d'aprenentatge supervisat. Treballen amb espais de dades d'altes dimensions (vectors p-dimensionals) que permeten trobar l'hiperplà òptim que defineix el conjunt de dades sota un concepte de marge.

En el cas de classificació, donat un conjunt de mostres d'entrada podem etiquetar les diferents classes que caracteritzen les dades i entrenar així una SVM per tal de construir un model que predigui la classe d'una nova mostra. És a dir, una SVM de classificació és un model que representa les mostres com a punts en l'espai i separa els grups de punts que pertanyen a la mateixa classe per l'espai buit més ample possible, aquest espai buit està definit per un hiperplà i la seva amplada s'anomena marge. D'aquesta manera, els punts del vector que són etiquetats amb una categoria estaran a un costat del hiperplà i els casos que es trobin en l'altra categoria estaran a l'altre costat.

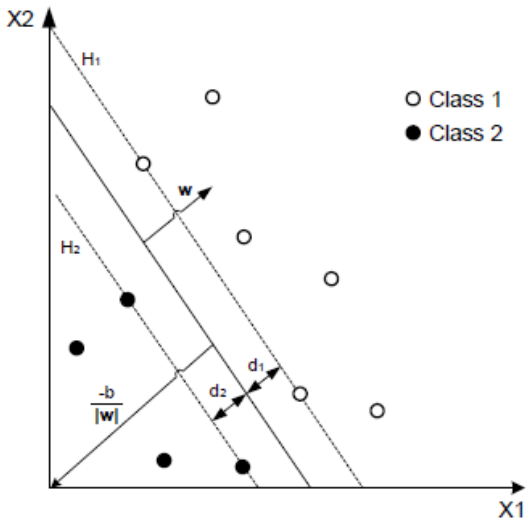


Fig.1 *Marge ampli de classificació binària lineal*. S. Sigh, 2005 [1]

En el cas de regressió, les màquines de vectors de suport construeixen un model també sota el concepte de marge. En aquest cas, però, quedaran dins del marge totes aquelles mostres desviades a menys d'una certa tolerància (ε) predefinida. És a dir, l'objectiu és trobar funció que generalitzi millor el conjunt d'observacions d'entrada sota una desviació màxima.

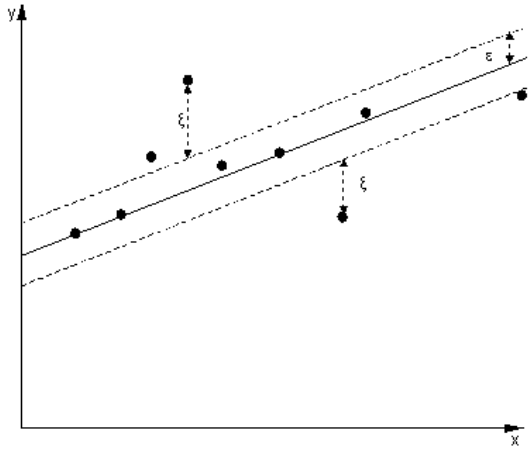


Fig.2. *SVM per a classificació. Font: kernelsvm.tripod.com*

És important remarcar que SVM és capaç de trobar fronteres no lineals gràcies a la funció de Kernel que veurem més endavant. També és una característica important que no només es pot aplicar la classificació per problemes binaris, sinó que també permet resoldre problemes multi-classe. Tot plegat suposa que SVM sigui una poderosa eina de classificació i regressió, però també una eina de complexitat matemàtica i computacional.

3.3.2 SVR: Regressió amb Màquines de Vectors de Suport

Per a establir un model predictiu de produccions energètiques horàries, cal una eina per construir models de regressió. En aquest projecte es proposa les màquines de vector de suport per a regressió com a eina predictiva per a aquestes series temporals. Cal comentar que els vectors de suport per a regressió (SVR) són una modificació de els vectors de suport per a classificació (SVC) que posteriorment es descriuran, però que en aquest treball ocupen un segon pla.

El mètode d'aprenentatge de suport vector per a regressió tracta d'estimar una funció òptima que aproximi les dades d'entrenament amb el menor error (ε), sota un concepte de marge.

És el que es coneix com tècnica de ε-SVR.

Sigui un problema genèric on tenim un conjunt d'exemples $\{ (x_i, y_i) \mid i \in \{1, \dots, n\}\ x_i, y_i \in \mathbb{R}^q \}$, on (x_i, y_i) són les mostres o vectors i q és la seva dimensió. La tècnica de ε-SVR busca una funció error (o de cost), anomenada ε-insensible, que sigui el més plana possible i en que els seus valors $\bar{y}_i$ difereixin de la sortida real y_i no més d' ε_i per totes les dades d'entrenament x_i. S'anomena vectors de suport a tots les observacions que delimiten els marges del model.

En el cas per a funcions lineals, es pot descriure la funció ε-insensible com:

$$f(\mathrm{x}) = (w^T x_i + b)\ \mathrm{w}, \mathrm{x} \in \mathbb{R}^q \text{ i } \mathrm{b} \in \mathbb{R}$$

Quan es diu que es busca una funció el més plana possible, significa que es busca una w petita (concepte de matemàtica euclidiana) que ens porta a la minimització de la seva norma:

$$\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2\ o\ \min_{w,b,\xi}\ \frac{1}{2} w^T w$$

subjecte a:

$$y_i - (w^T x_i + b) \leq \varepsilon_i,$$

$$(w^T x_i + b)\ -\ y_i \leq \varepsilon_i\ ,$$

No obstant, la majoria dels casos no hi ha solució convexa factible per a aquest problema i no es poden ajustar totes les parelles (y_i ,x_i) amb una precisió ε_i. En alguns casos cal tolerar alguns errors i deixar que algunes observacions quedin fora del marge. Per fer front a aquestes limitacions d'optimització s'introdueix el concepte de "marge suau" amb la incorporació d'unes variables de folgança ξ_i al problema d'optimització.

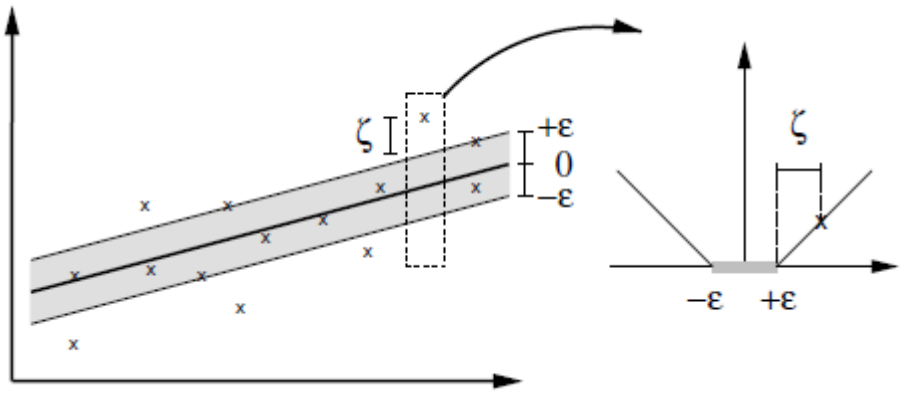


Fig. 3 *Ajust de marge suau per a un SVM lineal*. A. Smola and B. Scholkopf, 2002 [5]

Una dels principals dificultats que suposa SVM és l'ajust del marge, doncs en funció de la seva rigidesa o suavitat es definirà el model. Per tal de controlar la flexibilitat del marge, els SVM tenen un paràmetre C anomenat cost que controla la compensació entre errors d'entrenament i els marges rígids, per tal de definir un marge suau adequat. La identificació d'aquest paràmetre serà una de les tasques més importants a l'hora de modelitzar.

Ara per a trobar la funció òptima cal donar solució al següent problema d'optimització:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

subjecte a

$$y_i - (w^T x_i + b) \leq \varepsilon_i - \xi_i ,$$

$$(w^T x_i + b) - y_i \leq \varepsilon_i - \xi_i^* ,$$

$$\xi_i, \xi_i^* \geq 0.$$

On ξ_i és una variable de folgança i C és el paràmetre que la penalitza, de manera que quan més gran sigui C, més vectors hi haurà dins del marge i menys vectors de suport hi haurà. D'aquesta manera les mostres que estan a una distància del pla menor que ε no seran penalitzades ni tingudes en compte pel sistema. El paràmetre ε determina l'amplada de la zona ε-insensible que s'utilitza per ajustar les dades d'entrenament, ja que és el residu com a diferència entre els vectors de suport que delimiten el marge i les mostres que formen l'ajust del model [Fig.3].

Com a conclusió, l'estimació dels paràmetres d'una SVM és equivalent a la solució d'un

model de programació quadràtica amb restriccions lineals, amb la seva resolució pel mètode de multiplicadors de Lagrange, porta a trobar la solució òptima global i única (a diferència d'altres mètodes d'aprenentatge que només poden trobar mínims locals). No obstant cal anotar que, donat que aquest treball té per objectiu únicament la implementació de SVR i no la resolució dels algoritmes, la construcció i la resolució dels algorismes queden fora de l'abast del treball i s'utilitzaran més endavant, algorismes ja programats per experts.

3.3.2.1 Kernel

A la pràctica, no sempre les dades es distribueixen de manera que puguem aplicar una regressió lineal. Una solució a aquest problema és la representació mitjançant la funció de Kernel, que projecta la informació en un espai dimensional més gran, ampliant la capacitat de les màquines d'aprenentatge lineal. És el que es coneix com Mètodes Kernel.

Kernel SVR aplica transformació al seu conjunt de dades abans de l'etapa d'aprenentatge. Això li permet recollir les tendències no lineals en el conjunt de dades, a diferència de la regressió lineal.

La idea, representada a la Fig. 3, és transformar el conjunt de dades d'entrada X mitjançant $\emptyset(x)$, tal que $\emptyset : \mathbb{R}^{n+} \rightarrow \mathbb{R}^{n+}$, on n + > n, en un espai de dimensió alta on les dades es puguin modelitzar amb una recta. Aquesta transformació es pot expressar com el producte escalar dels vectors en l'espai de sortida:

$$K(x_i, x_j) = \emptyset(x_i) \cdot \emptyset(x_j)$$

El nou espai Z on es representen les dades transformades es coneix com espai d'Hilbert.

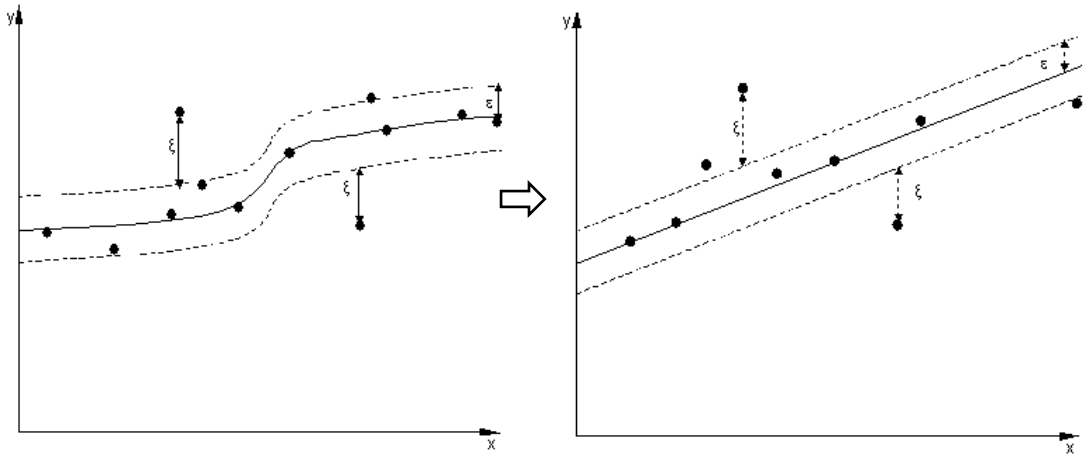


Fig. 4 Transformació de regressió mitjançant Kernel. Font: kernelsvm.tripod.com

Hi ha diferents tipus de mètodes Kernel. Els més bàsics i freqüentment utilitzats són:

Lineal: $K(x_i, x_j) = x_i \cdot x_j$

Polinomial: $K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d$, on $\gamma > 0$

Funció de base radial (RBF): $K(x_i, x_j) = \exp(-\gamma \cdot \|(x_i - x_j\|^2)$, on $\gamma > 0$

Sigmoide: $K(x_i, x_j) = tanh(\gamma x_i \cdot x_j + r)$

On $\gamma, r\ i\ d$ són paràmetres del Kernel.

Part del treball d'implementació serà escollir quina funció Kernel utilitzarem i quin valor atribuirem als seus paràmetres.

3.3.3 SCV: Classificació amb Màquines de Vectors de Suport

Sigui un problema genèric on tenim un conjunt d'exemples $X = \{x_i | i \in \{1, \dots, n\}\ ^\wedge\ x_i \in \mathbb{R}^q\}$, on q és la dimensió dels vectors d'entrada, i les seves etiquetes o classes $Y = \{y_i | i \in \{1, \dots, n\}\ ^\wedge\ y_i \in (1, -1)^l\}$ desconegudes. Aquests exemples s'utilitzen per entrenar el classificador per tal d'obtenir $\hat{y}_i = f(w, x_i)$, on w ∈ W, per proporcionar una etiqueta donada una entrada x_i. L'objectiu del procés d'aprenentatge és seleccionar de manera òptima la funció f(x, w), de manera que es minimitzin les discrepàncies entre les etiquetes vertaderes i les que proporciona la màquina de vectors suport.

3.3.3.1 Classificació Binària Lineal

En el cas dicotòmic o binari només hi ha dues classes, les classes poden ser etiquetades, per exemple, com $y_1 = 1$ per a la classe 1 i $y_2 = -1$ per a la classe 2.
En el cas bàsic on tenim dues variables predictores x_1 i x_2 –ens trobem a l'espai R^2 – i el conjunt de dades es linealment separable, tenim una representació plana com la de la Fig.5.

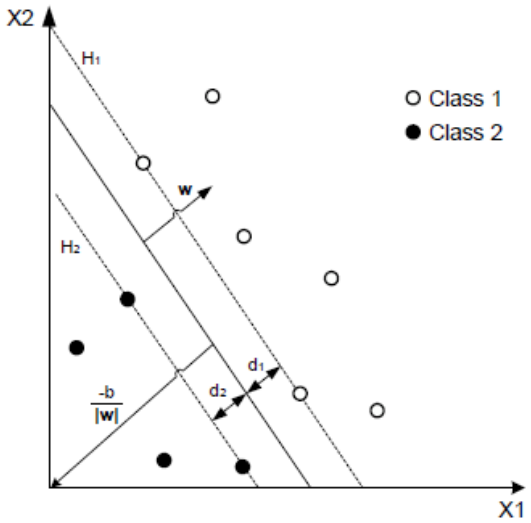


Fig. 5. *Marge ampli de classificació binària lineal*. S. Sigh, 2005 [1]

L'hiperplà de decisió que separa ambdues classes es pot descriure mitjançant l'equació:

$$\text{H:} \quad w \cdot x + b = 0$$

On $x \in R^q$ i $b \in R$. w és la normal al hiperplà i $\frac{b}{\|w\|}$ és la distancia perpendicular al hiperplà fins l'origen.

Si el problema és linealment separable, existeixen infinits hiperplans que separen els exemples d'entrenament. Com s'ha esmentat anteriorment, l'algorisme de classificació de SVM troba l'hiperplà separador H –en el cas binari lineal, una línia– que dista dels punts més propers de les dues classes per un marge el més ample possible. Aquest marge és el que es coneix com *hard margin* o marge màxim.

Per a trobar aquest hiperplà òptim H, primer l'algorisme imposa que el conjunt de dades d'entrenament satisfaci la condició:

$$y_i \cdot (w \cdot x_i + b) > 1$$

Així, la cerca del hiperplà es pot plantejar com el següent problema d'optimització:

$$\max_{w,b} \frac{2}{\|w\|}$$

$$\text{ó el que és el mateix:} \quad \min_{w,b} \frac{1}{2} \|w\|^2$$

$$\min_{w,b} \frac{1}{2} w^T w$$

Els hiperplans auxiliars H_1 i H_2 limiten la regió on no hi ha punts d'entrenament. Els exemples d'entrenament que estan en els hiperplans H_1: $w \cdot x + b = 1$ i H_1: $w \cdot x + b = -1$ s'anomenen vectors de suport x_{si} i són els punts més propers a l'hiperplà òptim. Mentre els vectors de suport no canviïn, el hiperplà no canviarà i serà independent de tots els altres exemples. Finalment, el marge d del SVM és calculat com a distancia entre H_1 i H_2 i el seu valor és de $\frac{2}{\|w\|}$.

Després de determinar l'hiperplà òptim de separació de classes mitjançant l'entrenament, ja podem predir les classes de noves dades x' evaluant el signe de $w \cdot x' + b$:

- Per a $w \cdot x' + b > 0$, x' es classifica com a classe 1, $y_1 = 1$
- Per a $w \cdot x' + b < 0$, x' es classifica com a classe 2, $y_2 = -1$
- Per a $w \cdot x' + b = 0$, x' es inclassificable

No obstant, fins i tot en els simples casos binaris de dues variables predictores les dades sovint no són linealment separables i el *hard margin* de SVM és irresoluble. En aquests casos es permet introduir una variable de folgança ξ_i. Aquest nou marge es el que es coneix com *soft margin* o marge suau. Per a aquest tipus de marge alguns exemples de les dades d'entrenament seran erròniament classificats.

Un dels principals dificultats que suposa SVM és l'ajust del marge, doncs en funció de la seva amplitud o suavitat definirà el model de classificació com a sota-ajustat o sobre-ajustat. Per tal de controlar la flexibilitat, els SVM tenen un paràmetre C anomenat *cost* que controla la compensació entre errors d'entrenament i els marges rígids, definint un marge suau adequat.

Finalment, podem dir que SVM requereix solució al problema d'optimització:

$$\min_{w,b,\xi} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

subjecte a $y_i(w^T x_i + b) \geq 1 - \xi_i$,

$$\xi_i \geq 0.$$

3.3.3.2 Classificació Binària No Lineal

Malauradament, no és habitual trobar aplicacions on es pugui realitzar la separació de classes de forma lineal. La realitat és que sovint ens trobem amb casos més de dos variables predictores, amb complexes corbes no lineals de separació –de fet, hiperplans d'altes dimensions- irrepresentables i difícils d'imaginar. A més l'algorisme SVM ha de tractar

amb conjunts de dades que no sempre poden ser completament separades. Tot plegat suposa una gran limitació computacional del aprenentatge lineal, que fa inviable el seu ús per la majoria d'aplicacions reals.

Novament, la solució a aquest problema és la representació mitjançant la funció de Kernel, que projecta la informació en un espai dimensional més gran, ampliant la capacitat de les màquines d'aprenentatge lineal. És el que es coneix com Mètodes Kernel.

La idea, representada a la Fig. 6, és transformar el conjunt de dades d'entrada X mitjançant $\emptyset(x)$, tal que $\emptyset : \mathbb{R}^{n+} \rightarrow \mathbb{R}^{n+}$, on n + > n, en un espai de dimensió alta on les dades siguin linealment separables. El nou espai Z és coneix com espai d'Hilbert. Aquesta transformació es pot expressar com el producte escalar dels vectors en l'espai de sortida:

$$K(x_i, x_j) = \emptyset(x_i) \cdot \emptyset(x_j)$$

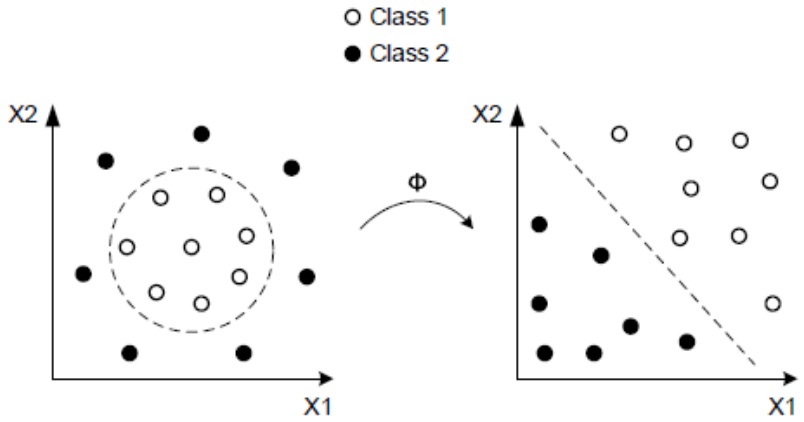


Fig. 6 *Transformació de dades mitjançant Kernel*. S. Sigh, 2005 [1]

3.3.3.3 Classificació Múltiple

El cas de la classificació de més de dos categories, $Y = \{y_i | i \in \{1, \dots, n\} \wedge y_i \in (1, -1)^k\}$, on k > 2, hi ha varies estratègies per resoldre-les, totes elles basades en una combinació de classificacions binàries que hem vist anteriorment.

Concretament, existeixen tres estratègies per resoldre SVC multi-classe:

- Mètode "Una Contra Totes" (*One versus All*, OvA): Técnica que k SVCs de forma independent. Cada classificador separa una classe respecte a la resta, determinant si una mostra pertany a aquesta classe. Davant d'un vector d'entrada

nou, s'avaluen totes les SVC binàries i se li assigna l'etiqueta corresponent a la que proporciona una major sortida.

- Mètode “Un Contra Un” (*One versus One*, OvO): S'entrenen $\frac{k\cdot(k-1)}{2}$ classificadors binaris, cadascun dels quals separa una parella de classes. Després, la decisió de la classe a la qual pertany una mostra nova es pot prendre convertint les sortides de les $\frac{k\cdot(k-1)}{2}$ SVCs binàries en k probabilitats a posteriori. Una altra opció molt habitual és emprar un mecanisme de votació, que assigna a la mostra l'etiqueta corresponent a la classe que hagi resultat vencedora en un major nombre de classificadors binaris. El mètode un contra un és el que s'utilitza en l'eina LibSVM que utilitzarem més endavant.
- Métode de “Graf Acíclic Dirigit” (*Directed Acyclic Graph SVM*, DAGSVM). Aquest graf té $\frac{k\cdot(k-1)}{2}$ nodes distribuïts en k - 1 nivells, amb un únic node en el nivell superior i k - 1 nodes en l'inferior. Cada un dels nodes consisteix en una SVC binària que compara una parella de classes. L'entrenament d'aquesta SVM multiclasse i els seus característiques fonamentals coincideixen, per tant, amb els del mètode un contra un. L'estructura en arbre i les decisions particulars que es prenen en cada nivell dirigeixen la decisió a través d'una branca del graf, produint la decisió final en l'últim nivell.

4. Implementació

En aquesta secció s'explica el procés d'implementació que s'ha seguit per a l'aplicació de les màquines de vectors de suport a les sèries temporals de produccions energètiques del parc eòlic La Espina. Primerament, es fa una revisió al conjunt de dades amb les que es treballa. Seguidament, es descriuen les eines informàtiques utilitzades i com emprar-les adequadament. Per últim, s'exposa el desenvolupament de les estratègies que s'han plantejat per a construir un sistema de predicció energètica basat màquines de vectors de suport per a regressió, arribant a uns resultats finals per al millor sistema obtingut.

4.1 Conjunt de dades

Les dades amb les que es realitza aquest estudi, són provinents del Parc Eòlic Espina, situat al municipi La Espina a la província de Lleó, Espanya, que consta de 9 aerogeneradors amb una potencia total instal·lada de 18kW.

Disposem de les dades de producció energètica horària total de 547 dies, 24 hores al dia. Concretament, tenim un arxiu de 24 columnes "N" corresponents a cada hora (on N=1,2,..,24 és la hora) de la fila "dia" (dia=1,2,...,547). Hem afegit una columna "Tot" que conté la suma de totes les produccions horàries per a cada dia. L'energia produïda per un aerogenerador s'expressa en unitats d'energia per unitat de temps, és a dir, potència. Aquesta potència és $P_v = \frac{A}{2} \cdot \rho \cdot v^3 \, [W]$, on A és l'àrea escombrada per les pales de l'aerogenerador, ρ és la densitat de l'aire i v la velocitat del vent.

Cal entendre que les sèries temporals relacionades amb energies renovables, i especialment amb energia eòlica, són molt difícils d'estudiar donat que estan estrictament relacionades amb factors meteorològics. Les dades meteorològiques són molt irregulars i caòtiques i això dificulta el seu estudi.

A més, la producció d'energia eòlica està condicionada per uns límits afegits per les característiques dels aerogeneradors. La corba habitual que descriu la relació entre la potència elèctrica lliurada per un aerogenerador i la velocitat del vent incident a la turbina ve representada a la Fig. 7, en el que es coneix com a corba de potència o corba P-V d'un aerogenerador.

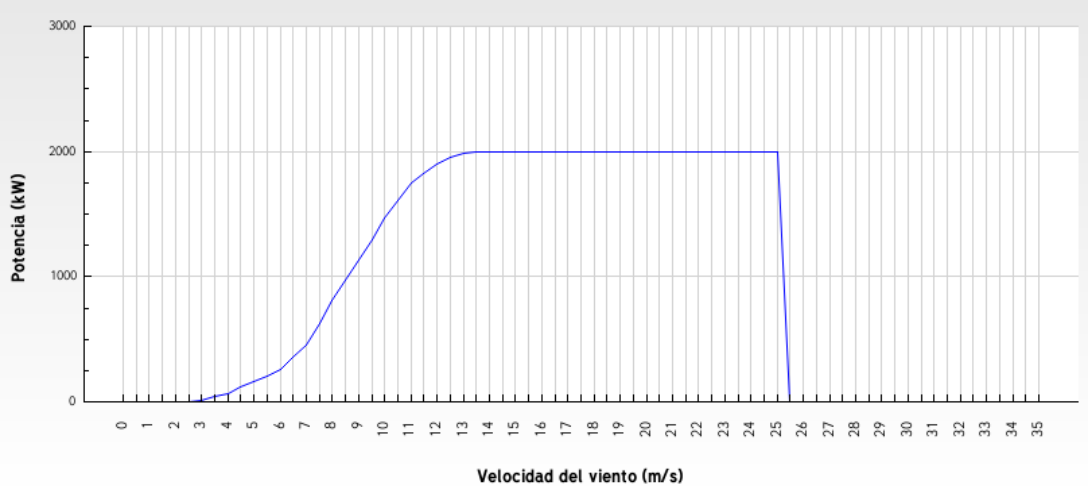


Fig. 7 Corba Potència – Velocitat del vent de l'aerogenerador Gamesa G87-2.0 MW, propi del parc eòlic La Espina. Font: www.thewindpower.net

Per a valors de velocitat de vent inferiors a 4 m/s, l'aerogenerador no produeix potència. Aquesta és l'anomenada velocitat d'arrencada de l'aerogenerador, a partir de la qual la potència generada presenta un creixement cúbic amb la velocitat del vent, fins a valors de vent d'uns 15 m/s. A partir d'aquesta velocitat del vent, la potència elèctrica generada es manté en el valor nominal per al qual va ser dissenyada la màquina (18kW en el nostre cas). Per a valors de velocitat de vent a partir d'uns 25 m/s, en el que es coneix com a velocitat de tall, es produeix la parada de l'aerogenerador per protegir enfront de vents massa severs.

En el punt de funcionament habitual, corresponent a velocitats de vent entre els 5 i 12 m / s, la corba de potència presenta un fort pendent, de manera que petites variacions de la velocitat del vent incident en el rotor provoquen grans variacions en la potència elèctrica generada, és a dir, la potència elèctrica és molt sensible als canvis de velocitat del vent. A causa d'això i al fet que el vent és una variable que pot tenir grans canvis en escales temporals molt curtes, les variacions en la potència lliurada per un aerogenerador al llarg del temps poden ser molt pronunciades.

4.1.1 Anàlisi exploratori de les dades

Representant les series temporals de cada hora (Fig.8), on l'eix vertical és la producció energètica i l'eix horitzontal és el temps en dies, podem veure que tenim dades molt irregulars i no es pot apreciar cap tendència ni estacionalitat a simple vista. S'observa que hi ha dies en que les produccions energètiques són nul·les o molt petites i al dia següent són molt grans.

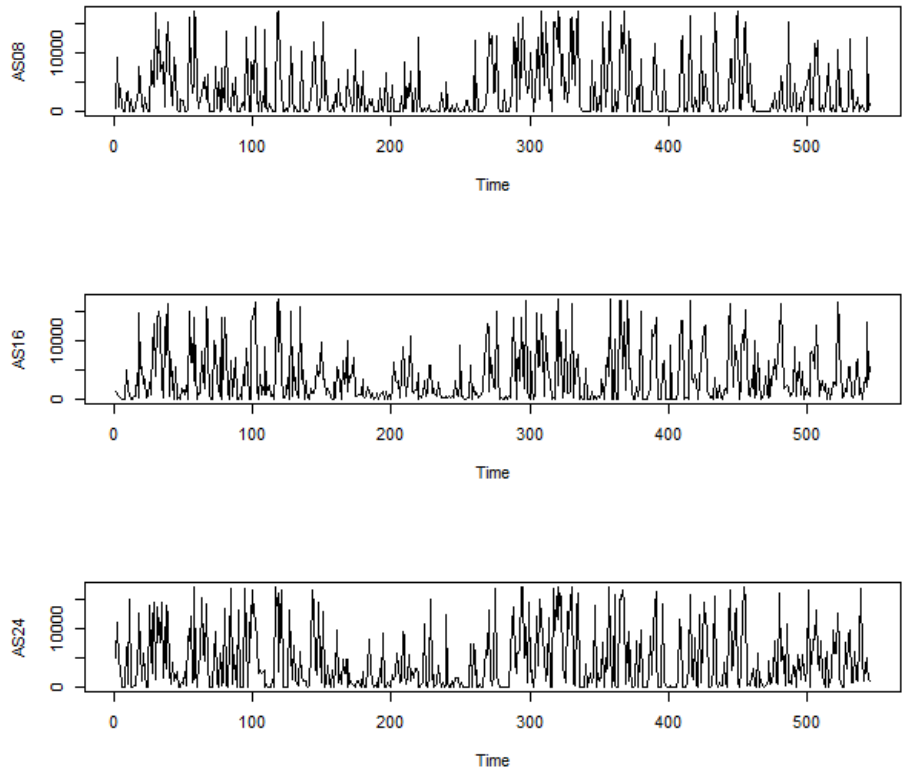


Fig. 8 Sèrie Temporal de produccions energètiques a les 8.00, 16.00 i 24.00.

Aquestes sèrie del conjunt de totes les produccions horàries dia a dia són molt difícil de modelitzar. El seu comportament caòtic causat per la irregularitat del vent com a factor meteorològic difícilment proporcionarà un bon model predictiu.

Representant les series temporals per hores cada dia Fig.9, és a dir, dia a dia les 24 hores del dia, tampoc s'aprecia cap tendència ni estacionalitat clara. Només prenent alguns dies consecutius ja veiem les sèries per a cada dia semblen ser totalment diferents.

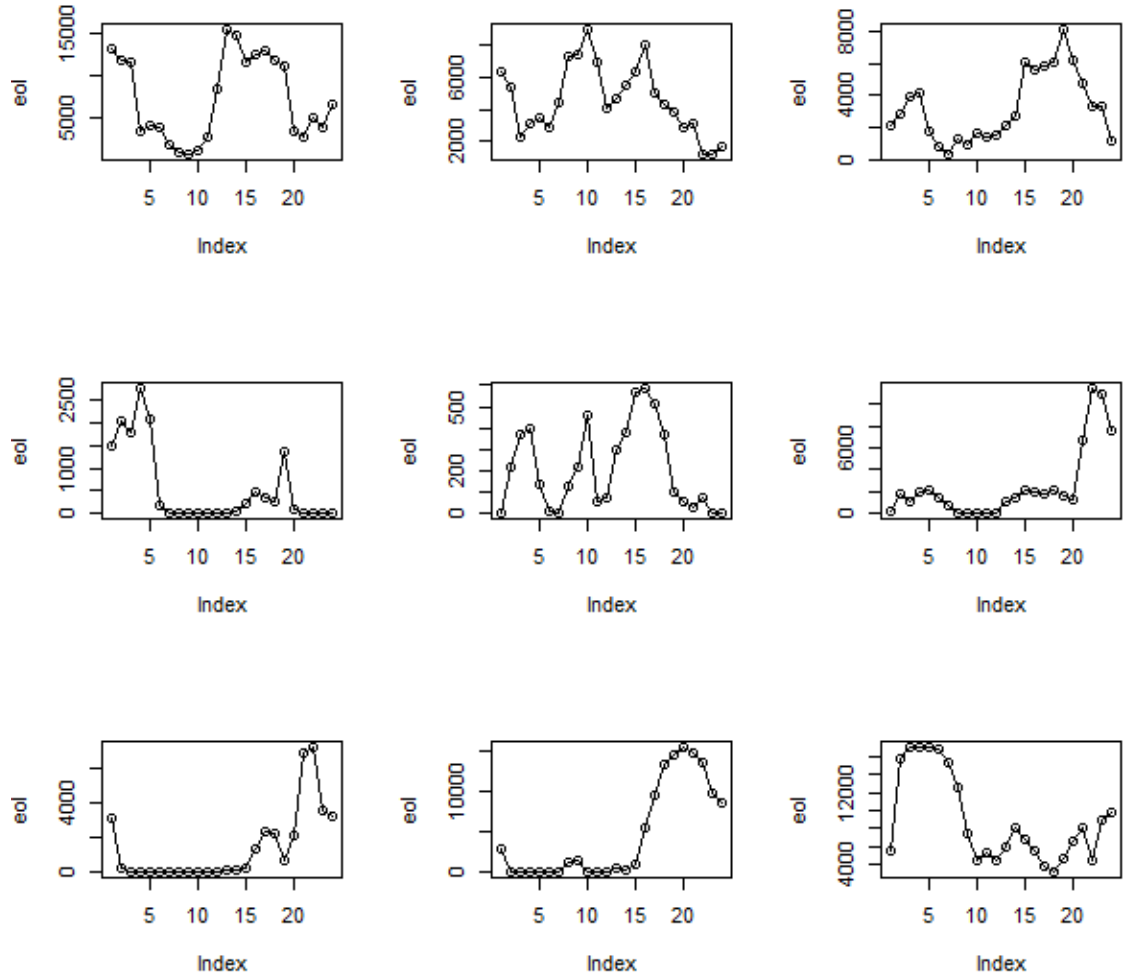


Fig. 9 Sèrie Temporal de produccions energètiques de 9 dies consecutius.

Veiem com les característiques d'engegada i tall de l'aerogenerador signifiquen directament que les nostres dades estiguin acotades entre 0 i 18 kW/h (potència total associada als 9 aerogeneradors del parc). Aquest és un tret característic important perquè la quantitat de 0 de les dades no es poca, ja que inclou les produccions tant dels dies de poc vent com les de massa vent, fet que suposa un increment addicional en la irregularitat de les dades. Si estudiem la sèrie temporal de la suma de produccions al dia (Fig. 10), veiem com no són pocs els dies on la producció predominant és la nul·la. El motiu és desconegut i a la vegada rellevant, per tant no podem prescindir d'aquestes dades com si fossin anòmales, sinó que determinar quan succeiran també és objecte de l'estudi.

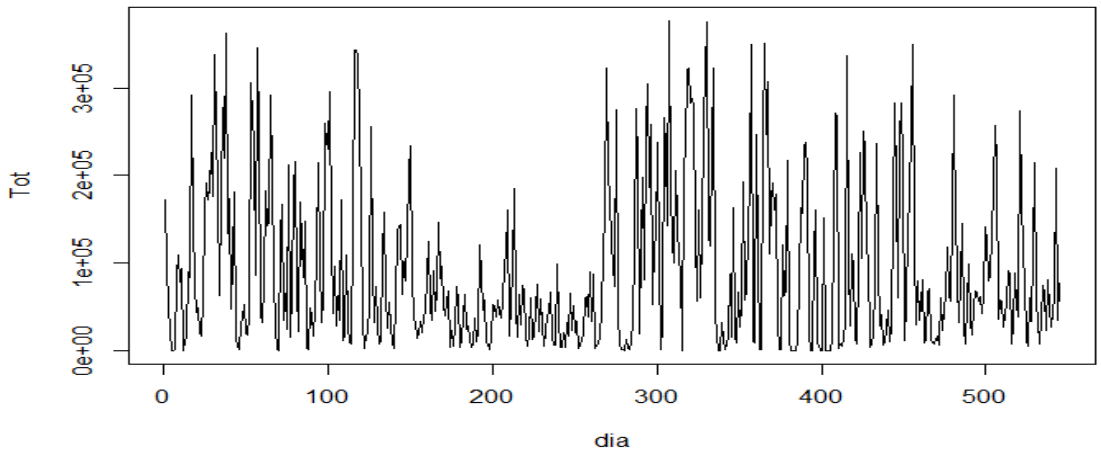


Fig. 10 Sèrie Temporal de la suma de produccions energètiques horàries.

A tall de conclusió, la irregularitat de les dades fa que l'estudi per els mètodes de series temporals clàssics resulti molt complex i s'espera que mitjançant el mètode d'aprenentatge supervisat SVM es pugui establir un model vàlid que ens sigui més útil per a realitzar previsions futures.

4.2 Aplicació de Suport Vector Machine

Els mètodes de SVM per regressió han estat aplicats a predicció de sèries temporals amb resultats satisfactoris en molts casos, però també cal tenir en compte que presenten certes dificultats importants. Les constants o paràmetres són dependents de cada problema particular i no hi ha mètodes que permeten la seva estimació, pel que els seus valors han de ser fixats per l'expert. Així, s'haurà de tenir en compte que:

1. No hi ha cap mètode per seleccionar la funció de nucli òptima per a un problema concret.

2. L'optimització dels paràmetres és costosa computacionalment. El procés d'optimització requereix de l'avaluació d'algun procediment de validació creuada o algun criteri bayesià amb complexitat elevada

3. Els models generats poden ser grans, ja que solen incloure grans conjunts de dades.

Per a fer front a aquestes dificultats i poder definir un model amb resultats acceptables, els reconeguts desenvolupadors de maquines de vectors de suport Chih-Wei Hsu, Chih-Chung Chang i Chih-Jen Lin [6] ens proposen seguir el procés seguent:

1. Adaptar el conjunt de dades per poder-les processar amb svm()
2. Escalar les dades
3. Plantejar quin mètode Kernel utilitzar
4. Utilitzar validació creuada per trobar els millors paràmetres C i γ

5. Entrenar utilitzant els paràmetres trobats
6. Testejar

4.2.1 Per què R?

Per a la implementació de les tècniques de SVM utilitzarem el software gratuït R-Studio, ja que té un llenguatge estadístic entenedor i és una eina potent que ens serà molt útil. R inclou alguns paquets específics de Machine Learning que ens permetran analitzar i visualitzar les dades. Això simplifica molt l'ús de SVM, ja que només ens caldrà conèixer bé com utilitzar els algorismes pertinents per poder solucionar el nostre problema.

El codi RScript utilitzat per a la implementació del treball s'exposa als annexos.

4.2.1.1 El paquet e1071

El paquet e1071 de R ofereix una interfície a la implementació en C++ *libsvm* de Chih-Chung Chang i Chih-Jen Lin. Aquest paquet inclou: mètodes de classificació, regressió i detecció d'anomalies basats en SVM, a més de funcions Kernel i validació k-creuada.

Una extensió que ens interessa especialment és la *ε-regression*, que aplica els fonaments matemàtics de ε-SVR descrits anteriorment per a construir un model predictiu basat en la maximització del marge.

Un tret característic important en l'ús del paquet en R, és que s'utilitza el mateix objecte *svm()* tant per classificació com per regressió. Si la resposta és una factor, ella mateixa canvia al mode classificació; en cas contrari, resta en mode regressió.

L'objecte que ens proporciona e1071 s'anomena svm() i es caracteritza pels atributs següents:

```
svm(x, y = NULL, scale = TRUE, type = NULL, kernel = "radial", degree
= 3, gamma = if (is.vector(x)) 1 else 1 / ncol(x),coef0 = 0, cost = 1,
nu = 0.5, class.weights = NULL, cachesize = 40, tolerance = 0.001,
epsilon = 0.1,shrinking = TRUE, cross = 0, probability = FALSE, fitted
= TRUE,..., subset, na.action = na.omit)
```

On:

- x són les dades d'entrada, en el nostre cas produccions energètiques del dia D-1.
- y és la resposta, en el nostre cas produccions energètiques del dia D.
- type és el tipus de mètode svm, en el nostre cas regressió epsilon insensible.
- gamma és el paràmetre del Kernel RBF.
- cost és el paràmetre de penalització del model.

- epsilon és el marge del model.
- cross indica la k de la validació creuada k vegades, si s'escau.

4.3 Models SVR

Com a primera estratègia per construir un sistema de predicció, es planteja utilitzar únicament les màquines de vectors de suport per a regressió. La idea és crear un model estocàstic per a cada hora dels dies D alimentats de la les dades de les 24 hores del dies D-1. Si aquest conjunt de dades (X,Y) del dia D-1 no aporta suficient precisió, es plantejarà utilitzar també produccions de dies anteriors (D-2, D-3,...).

4.3.1 Preprocessament de les dades

4.3.1.1 Estructuració de les dades

Primerament, s'adapta el conjunt de dades per a poder crear el model. S'ha decidit organitzar les dades per tal de realitzar un conjunt de 24 models de regressió de vectors de suport que determinin la producció de cada hora del dia D a partir de la producció de les 24 hores del dia anterior D-1. Així, les dades d'entrada seran les dades de produccions del dia D-1, i la resposta y_h de sortida seran les dades del dia D. Hi haurà un model per cada hora h=1,2,3,...,24 del dia D. D'aquesta manera s'espera poder construir una eina ens permeti realitzar bones prediccions a curt termini (h petites) i poder actualitzar les previsions intrahoràries –que són tan importants en el mercat elèctric- amb garantia, tot i que s'espera que per a hores més llunyanes (h grans) doni pitjors resultats.

A causa de que les dades meteorològiques siguin tan caòtiques, el conjunt de observacions presenta molta variabilitat. Per facilitar l'estudi s'ha realitzat una transformació logarítmica.

Per a la construcció dels models, primerament es procedeix a dividir el conjunt de dades dels 545 dies que tenim en dues parts: 445 per entrenar el model i 100 per testejarlo.

4.3.1.2 Escalament de les dades

Escalar les dades abans d'aplicar SVM és molt important. El principal avantatge és evitar que els atributs amb majors variàncies predominin respecte els de petites variàncies, aconseguint que la variància sigui més o menys constant. Un altre avantatge és evitar dificultats numèriques durant el càlcul, reduint així el cost computacional.

Per sort, la funció *svm*() del paquet e1071 ens proporciona automàticament un escalament de mitjana 0 i variància 1, mentre tinguem l'argument *scale* amb valor TRUE.

4.3.1.3 Selecció del model

Seleccionar el model consisteix en primer lloc seleccionar el Kernel més adequat. Seguidament escollir els millors paràmetres del Kernel i paràmetre C. Per últim, per ser un cas de regressió en que utilitzem el mètode ε-SVR, cal seleccionar la ε de la funció insensible.

Selecció del Kernel

Com hem vist a la revisió teòrica dels models de vectors de suport, hi ha diversos mètodes Kernel per a transformar les dades d'entrada. En general, el Kernel de funció de base radial o RBF ("*radial basis function*") és una primera elecció raonable per diversos motius:

- Pot treballar quan el model teòric no és lineal.
- És un Kernel genèric, ja que els Kernels lineal i sigmoide es comporten com un RBF per a certs paràmetres.
- Només afegeix un paràmetre al model (γ), mentre que els Kernels sigmoide i polinòmic resulten més complexos per estar caracteritzats de dos i tres paràmetres afegits respectivament.
- Numèricament és simple, ja que pot prendre valors entre 0 i 1. En canvi, el Kernel polinòmic pot prendre valor infinit o zero, i aquest fet ens podria suposar problemes.
- És el més utilitzat i sol donar resultats adequats, per això la funció svm() el pren per defecte.

En motiu de les seves múltiples avantatges s'ha decidit partir d'aquest Kernel i provar-ne d'altres en cas de que els resultats d'aquest no siguin prou bons. Als annexos [10.2] es pot trobar la gràfica comparativa del model SVR-lineal amb el model lineal.

Metodologia per a selecció de paràmetres

Donat que no hi ha una manera d'escollir els paràmetres, l'expert ha de determinar-los per poder crear el model. La metodologia utilitzada per a la selecció de paràmetres consisteix en introduir uns valors inicials per (γ, C,ε) i procedir a estimar els valors òptims mitjançant la comparació d'un estadístic que mesuri l'error d'ajust del model a la sèrie de temps.

En principi qualsevol tècnica d'optimització podria ser utilitzada per estimar els valors òptims del tres paràmetres, però l'obtenció de la solució de problema de programació quadràtica és computacionalment costosa. Una bona estratègia pot ser la recerca discreta a on s'avaluen seqüencialment els punts en el veïnatge de (γ, C,ε).

Una bona eina de recerca discreta per a trobar els paràmetres del model de regressió, és realitzar un "grid-search" de dos dels paràmetres a trobar (paràmetres més influents en el model), generant totes les possibles combinacions de valors per tal de trobar la millor. Després d'identificar els valors dels paràmetres, podrem construir el model final amb totes les dades d'entrenament. El paquet e1071 inclou un objecte *tune.svm()* que ens facilita aquesta feina, realitzant una gran quantitat de models per a les diferents parelles de paràmetres, i escolleix la que té un RMSE més baix,

on RMSE=$\sqrt{\frac{\sum(Y-\hat{Y})^2}{n}}$.

L'objecte tune.svm() és el següent:

```
tune(method, train.x, train.y = NULL, data = list(), validation.x =
NULL, validation.y = NULL, ranges = NULL, predict.func = predict,
tunecontrol = tune.control(), ...)
```

On:

- x són les dades d'entrada, en el nostre cas produccions energètiques del dia D-1.
- y és la resposta, en el nostre cas produccions energètiques del dia D.
- gamma és el paràmetre del Kernel RBF.
- cost és el paràmetre de penalització del model.
- epsilon és el marge del model.

Aquest objecte assigna automàticament els millors valors a unes constants que les criderem amb *"tuned$best.parameters$".*

Selecció de paràmetres (γ, C,ε)

Una vegada seleccionada la funció Kernel, cal identificar els millors paràmetres que ajusten un model SVR acurat per a poder predir correctament dades desconegudes.

En el cas d'utilitzar el Kernel RBF, és important determinar el millor paràmetre gamma propi d'aquest Kernel. Aquet paràmetre es caracteritza per constituir models complexes quan pren valors grans, i models més simples per a valors baixos. Per a valors molt grans, doncs, sobre-ajusta el model, cosa que no ens convé per fer prediccions futures, però per valors massa petits no ajusta bé les dades, les sota-ajusta. De fet, per a un determinat valor petit realitza una mapeig molt semblat al que realitza el Kernel lineal.

El paràmetre ε controla defineix directament l'amplada de la zona ε-insensible i per tant també és un paràmetre important a "tunejar". Pren valors entre 0 i un valor positiu desconegut que representa l'error màxim entre el model i la sèrie de dades.

El paràmetre C és el paràmetre que regula la suavitat del marge, és a dir, que penalitza les mostres que queden fora del model. Així, quant més gran sigui el cost menor quantitat de vectors de suport, cosa que incrementa la precisió del model. Per contra, massa precisió pot significar sobre-ajust, per tant cal anar amb compte amb el valor que assignem a C.

Per a trobar el valor adient d'aquest paràmetres s'utilitza la metodologia clàssica i recomanada per Chih-Wei Hsu, Chih-Chung Chang i Chih-Jen Lin [6]. Primerament, es realitza una exploració dels paràmetres per tenir una idea de com afecten al model. Partint d'aquest valors relativament bons, s'explora la millor combinació de valors propers als trobats per tal de trobar el millor model. Aquest mètode d'exploració és el comunament conegut com "grid-search".

Per avaluar la efectivitat d'aquests paràmetres es compara, mitjançant l'estadístic $RMSPE = \sqrt{\frac{\sum(Y-\hat{Y})^2}{Y}}$, el model obtingut de regressió de vectors de suport amb un model de regressió lineal clàssic, provant diferents valors de paràmetres. Es recorda que el model lineal es caracteritza per ser un model paramètric que ajusta la resposta en funció de les respostes anteriors de la sèrie.

Cal tenir en compte que donat que per construir el model es prenen dades conegudes, es corre el risc de sobre ajustar el problema a les dades, fet que implicaria una molt bona predicció en les dades d'entrenament però una mala predicció de dades futures desconegudes. Per evitar aquest problema, una bon mètode comunament utilitzat és el *K-fold cross validation*. Consisteix en dividir les dades d'entrenament en k subconjunts d'igual mida i, seqüencialment, es testeja un subconjunt utilitzant el model entrenat per k-1 subconjunt i es va canviant el subconjunt de test fins a haver realitzat un total de k entrenaments i k tests. Així cada mostra és predita i s'obté un percentatge d'encert de predicció. D'aquest manera es pot detectar prèviament al testeig, si s'està sobre-ajustant.

D'aquesta manera una bona estratègia per ajustar el model es realitzant una recerca dels paràmetres mitjançant un "grid-search" a la vegada que s'utilitza una validació creuada. La funció svm() del paquet e1071 proporciona la opció de k-fold cross validation. Només caldrà indicar la k desitjada al argument *cross=k*. Si no s'indica cap valor, svm() pren k=0. La funció ens retornarà els MSE resultant de cada encreuament per a que puguin ser seguidament analitzats.

1) <u>Exploració del paràmetre C:</u>

Realitzant una primera exploració del paràmetre C (deixant els paràmetres γ i ε fixos), amb les dades que s'utilitzaran d'entrenament, els resultats obtinguts han estat:

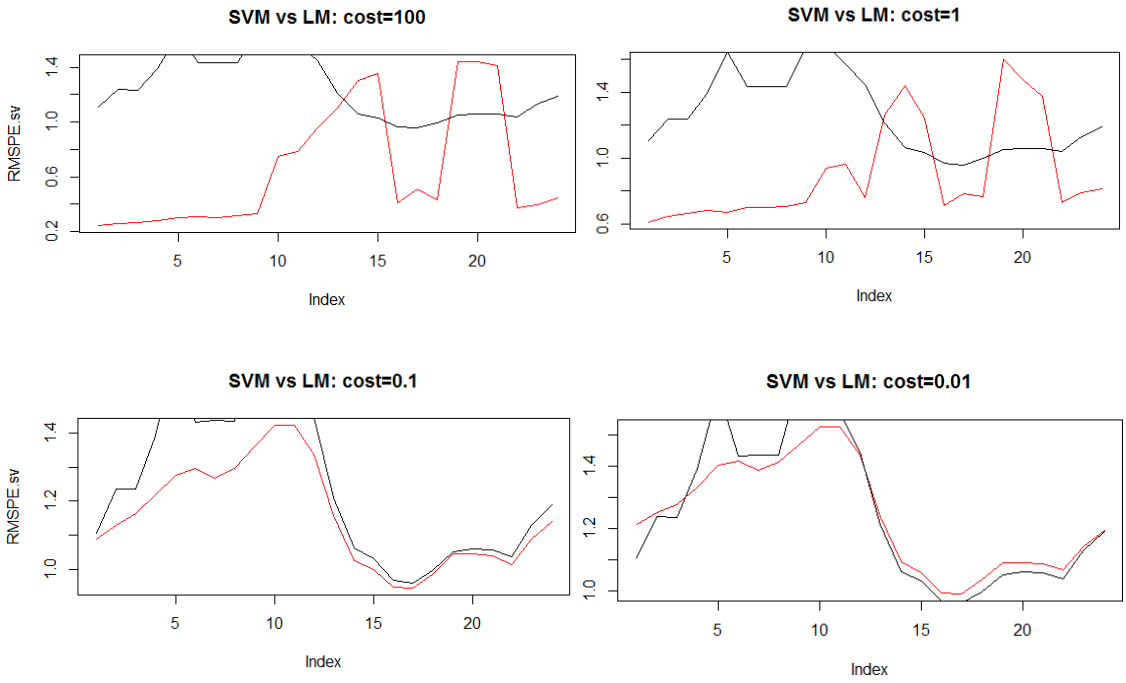
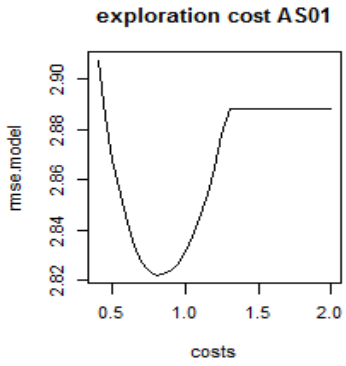
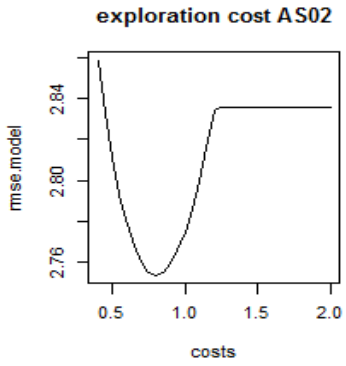
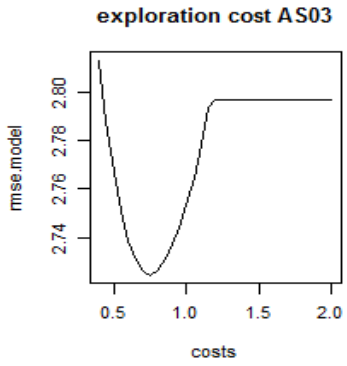


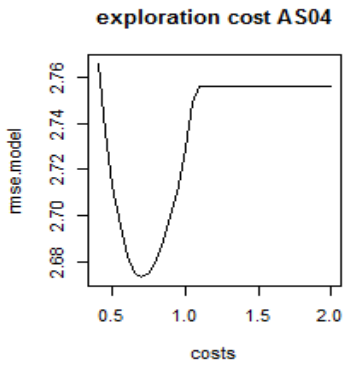
Fig. 11 *Comparativa del RMSPE resultant del models SVR (vermell) i lineal (negre) sota valors diferents de cost.*

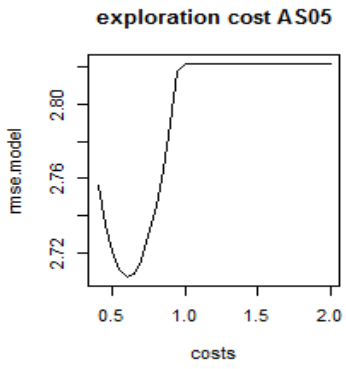
El resultat ha estat que els valors dels residus per a les primeres hores són millors amb un cost elevat, mentre que els residus a les últimes hores són millors amb un cost baix. Cal estudiar amb més profunditat aquest comportament.

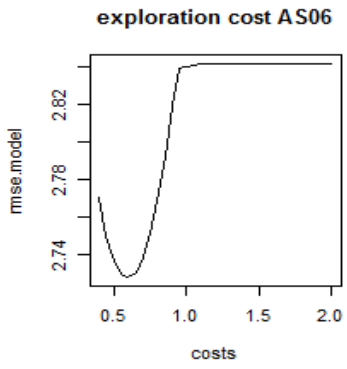

exploration cost AS01
rmse.model
costs

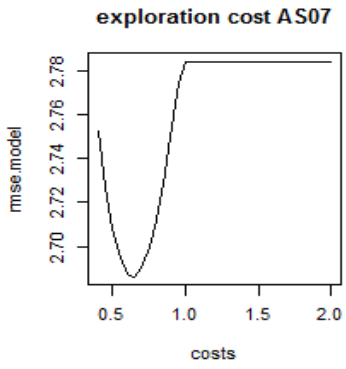

exploration cost AS02
rmse.model
costs

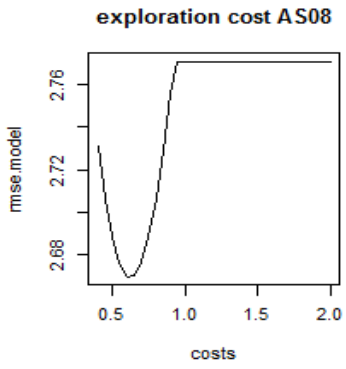

exploration cost AS03
rmse.model
costs

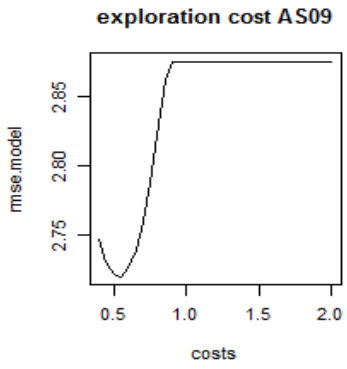

exploration cost AS04
rmse.model
costs

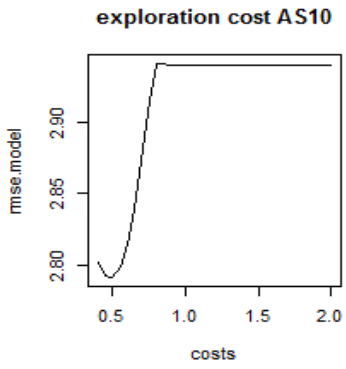

exploration cost AS05
rmse.model
costs

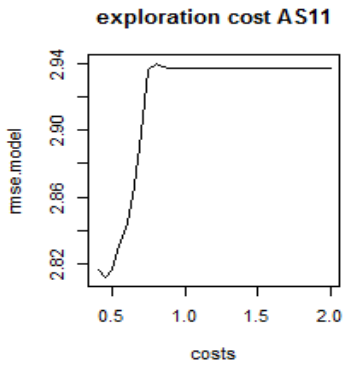

exploration cost AS06
rmse.model
costs

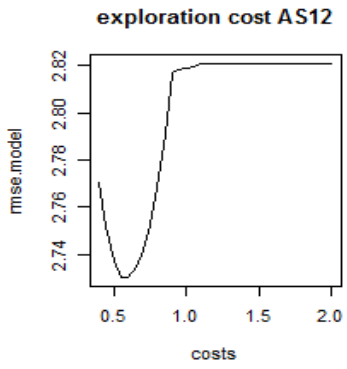

exploration cost AS07
rmse.model
costs


exploration cost AS08
rmse.model
costs


exploration cost AS09
rmse.model
costs


exploration cost AS10
rmse.model
costs


exploration cost AS11
rmse.model
costs


exploration cost AS12
rmse.model
costs

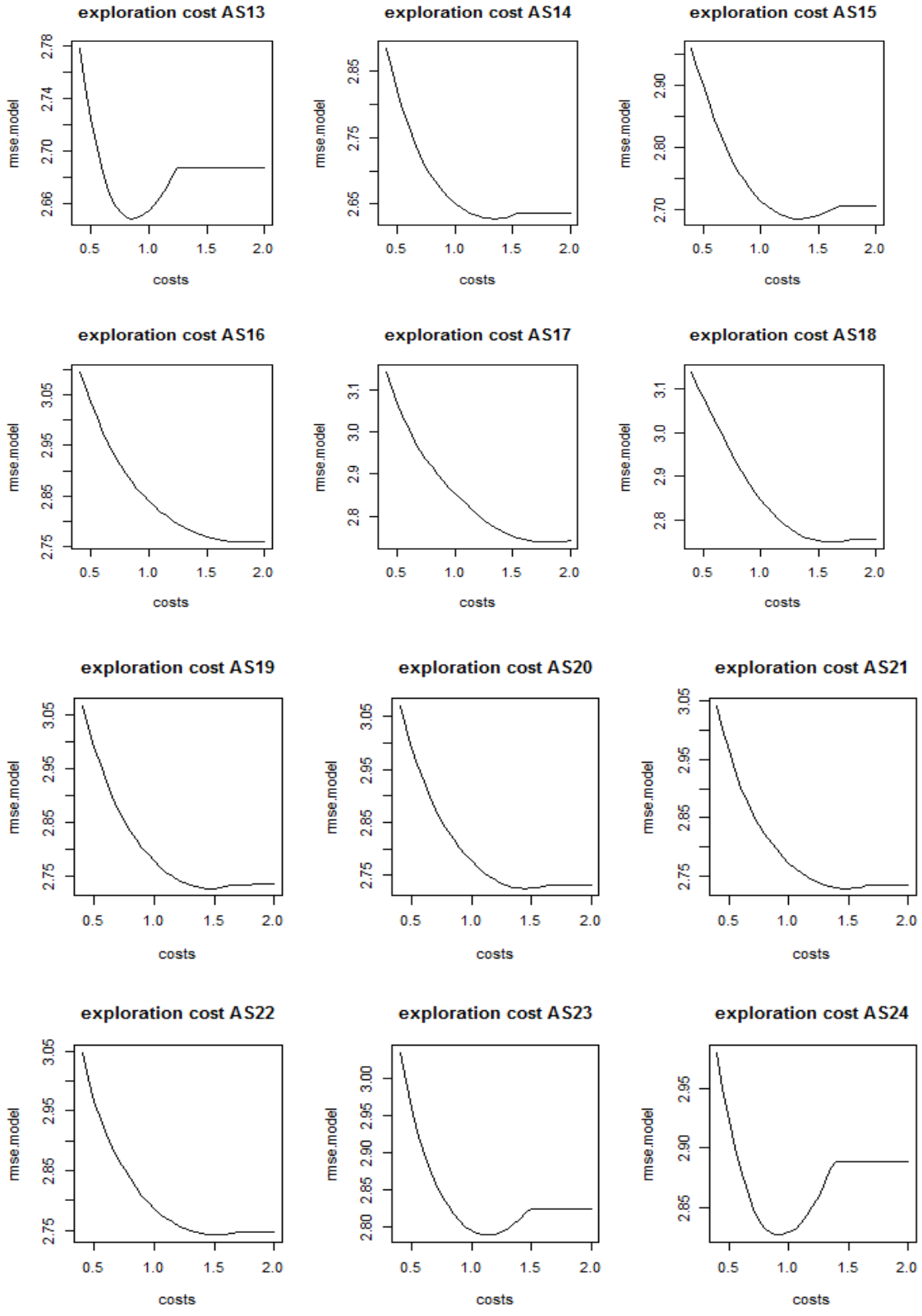


Fig. 12 Comparativa del MSE resultant de cada model SVR sota valors diferents de cost.

Avaluant el RMSE a diferents costos per a diferents models horaris, però realment estan acotats entre 0.4 i 2. Es veu com els millors costos per a les primeres hores (hores futures properes a les dades d'entrenament), els valors òptims oscil·len entre 0.5 i 1. Després entre les hores 13 i 22 els costos òptims són alts. En canvi, per a models de les ultimes dues hores del dia (hores llunyanes a les dades d'entrenament), els valors dels millors costos tornen a ser petits.

Cal afegir que interessa que els costos siguin baixos per tal de no tenir models sobreajustats, i el fet de que el millor resultat sigui per valors al voltant de 1 (i no al voltant de 100 com podria haver passat) és satisfactori.

La variació dels valors dels millors costos ens confirma que ajustar un model diferent per a cada hora és una bona estratègia de modelització en aquestes dades. No obstant, tot i que els millors models s'ajusten amb costos diferents, una bona idea per generalitzar-los tots pot ser ajustar tots els models amb un mateix cost prou bo. Observant en els gràfics, es veu com per a costos de 2 tots els residus són força bons i el cost es prou baix. Alhora de fer la decisió final dels models es plantejarà quina de les dues estratègies ens convé més.

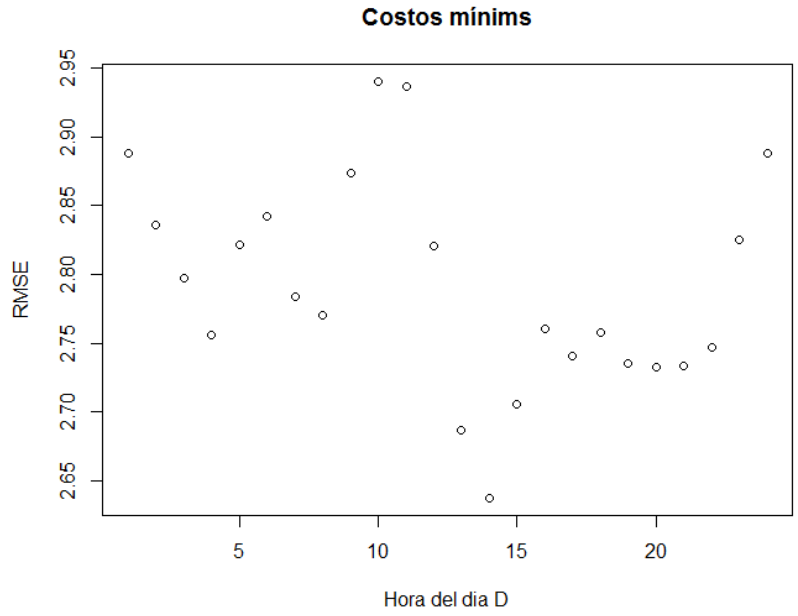


Fig. 13 Comparativa del RMSE resultant de cada model SVR per al millor valor de cost respectiu.

Com podem veure a la figura, els residus obtinguts per a els millors models de cada hora són satisfactòriament baixos.

2) Exploració del paràmetre ε:

Realitzant una primera exploració del paràmetre ε (deixant els paràmetres γ i C fixos), amb les dades que s'utilitzaran d'entrenament, els resultats obtinguts han estat:

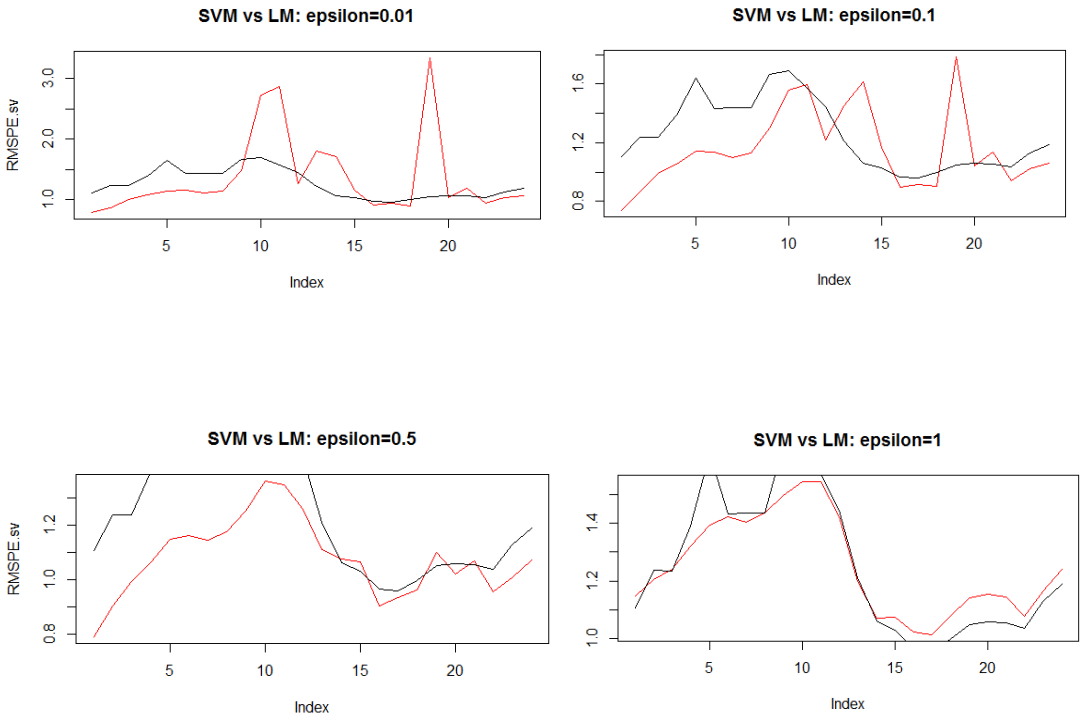


Fig.8

Fig. 14 Comparativa del RMSPE resultant del models SVR (vermell) i lineal (negre) sota valors diferents de ε. Font pròpia RStudio

Com es pot observar a la Fig. 8, el resultat ha estat que els valors dels residus són millors per a un valor de ε aproximat a 0.5.

3) Exploració del paràmetre γ:

Realitzant una primera exploració del paràmetre γ (deixant els paràmetres ε i C fixos), amb les dades que s'utilitzaran d'entrenament, els resultats obtinguts han estat:

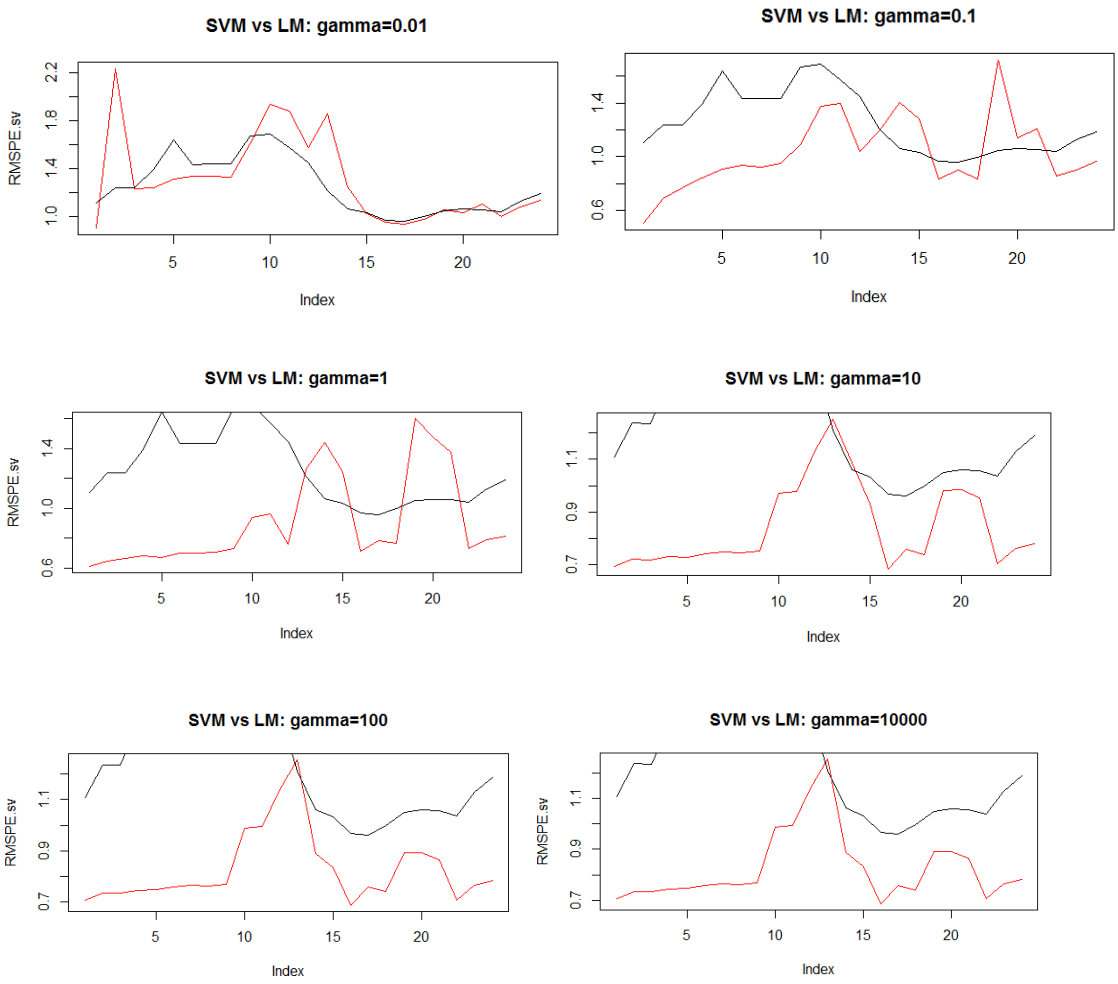


Fig. 15 Comparativa del RMSPE resultant del models SVR (vermell) i lineal (negre) sota valors diferents de γ

S'observa que per a valors més grans de 10 el model és considerablement millor, però quan a partir de 100 ja no hi ha quasi diferències i el model no canvia més. Valors molt grans de gamma creant models més complexos i sovint sobreajustats. Un valor de gamma 10 sembla prou bo.

Finalment, probem l'entrenament per als millors valors obtinguts:

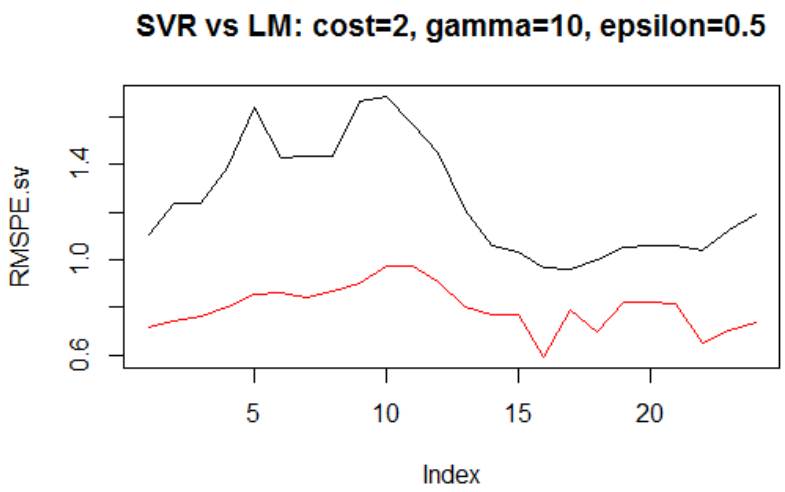


Fig. 16 Comparativa del *RMSPE resultant del models SVR (vermell) i lineal (negre) amb els millors paràmetres trobats manualment. Font pròpia RStudio*

El resultat mostra que clarament el model de regressió per vectors de suport és molt millor que l'obtingut amb el model de regressió lineal quan fem prediccions sobre les dades d'entrenament. Posteriorment caldrà realitzar prediccions sobre dades de testeig, no d'entrenament, per validar que realment sigui bon predictor.

Un cop trobats uns bons valors de partida, es precedeix a utilitzar la funció *tune()* per a seleccionar de forma automàtica els millors paràmetres (respecte els pertanyents MSE). S'utilitza doncs una seqüencia de valors al voltant dels valors de partida. Concretament els valors han estat: un cost entre 0.1 i 2 amb salts de 0.1, una epsilon entre 0.2,0.8 amb salts de 0.1, i una gamma constant de 10. El resultat ha estat:

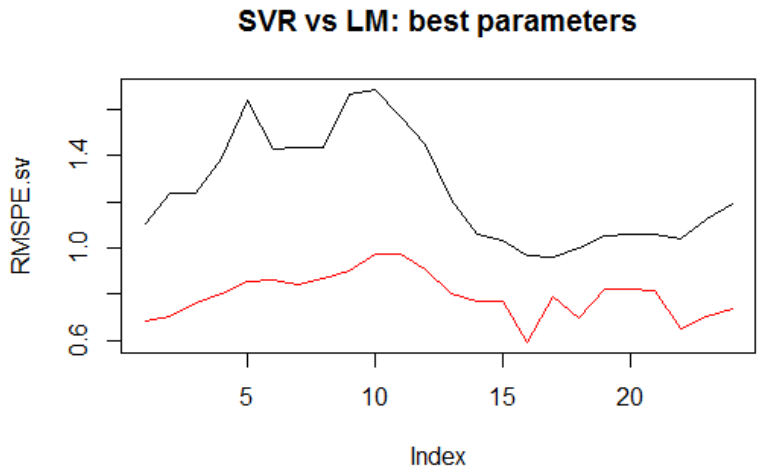
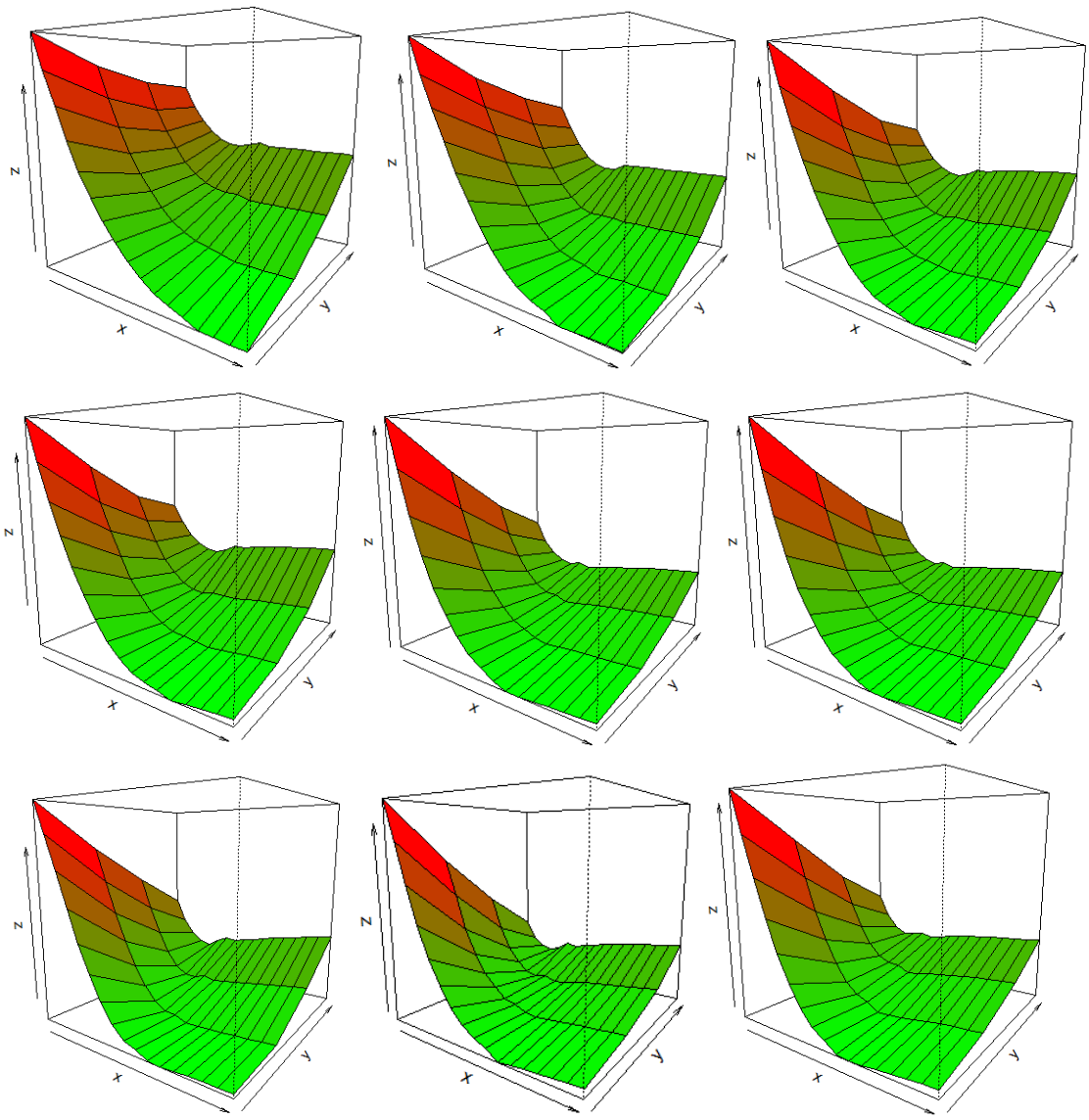
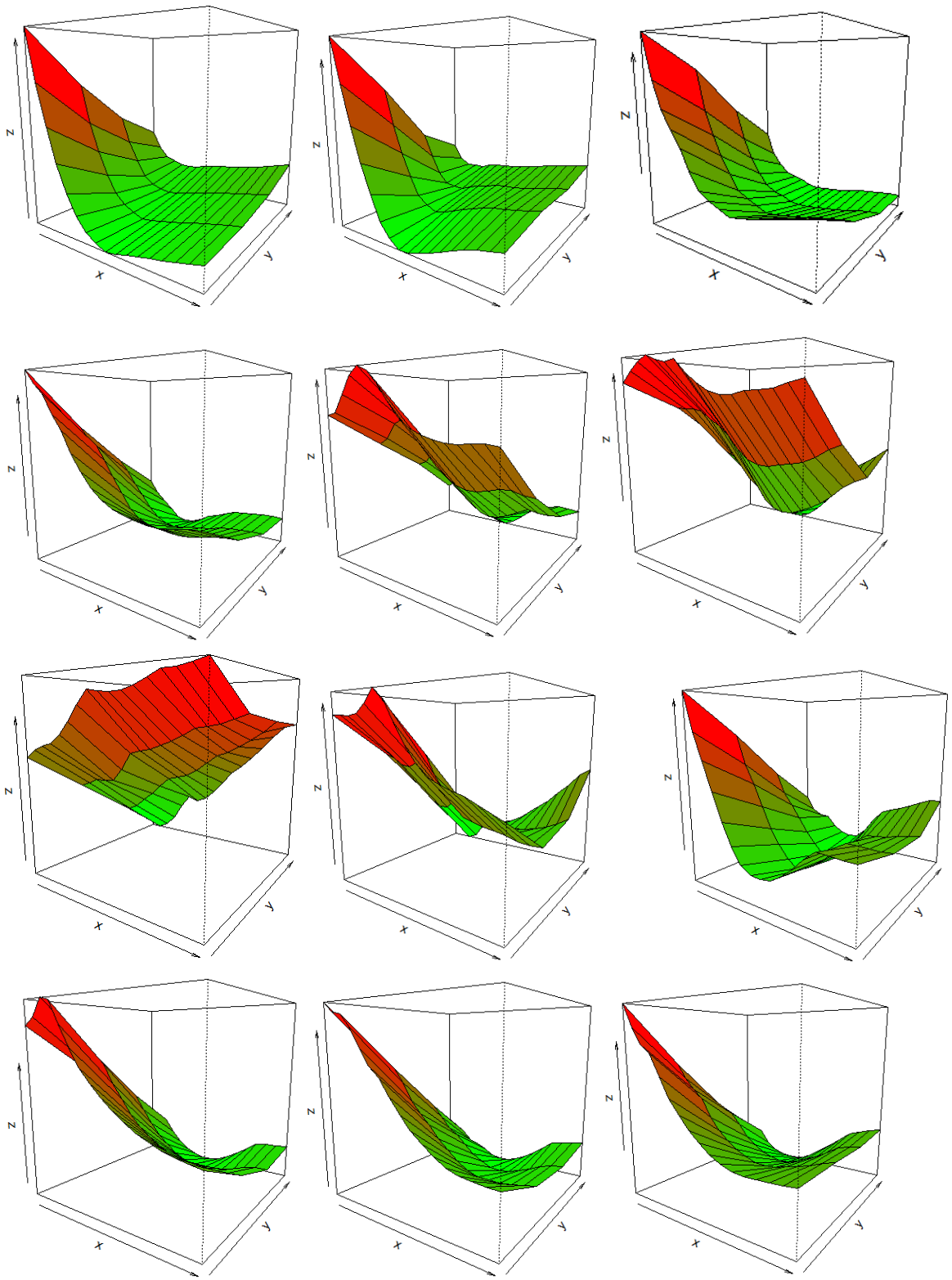


Fig. 17 Comparativa del *RMSPE resultant del models SVR i lineal amb els millors paràmetres trobats mitjançant "grid-search". Font pròpia RStudio*

Sorprenentment, els residus no han estat millors que els trobats manualment per totes les hores, sino aproximadament els mateixos. Això pot ser degut a que haguem trobat els millors valors en la primera exploració.

Per tal de veure com treballa l'objecte tuned(), s'han graficat el resultat en una representació 3D:




z
x
y

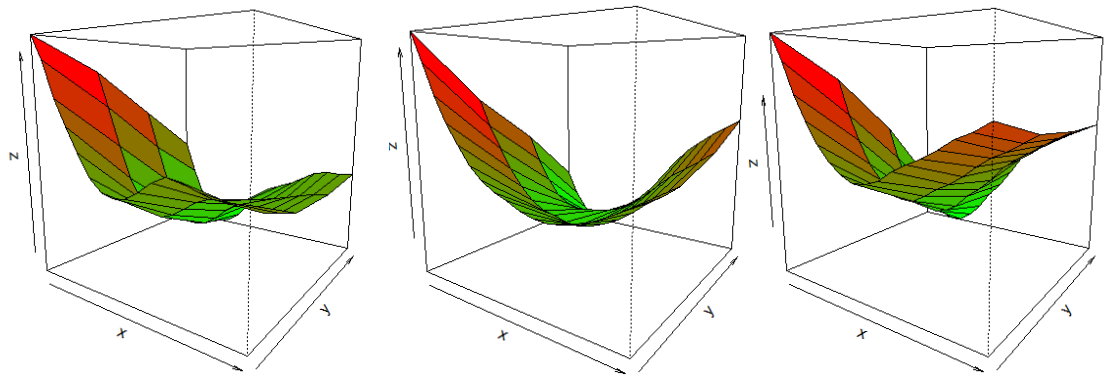


Fig. 18 Comparativa de cada model SVR, on les x és el cost, y és epsilon, z és l'error mesurat com MSE.

Com podem veure, per a cada model (cada hora) els paràmetres òptims són diferents. L'objecte tuned permet detectar aquests millors paràmetres que després són implementats a l'objecte svm().

Per a verificar que el millor model sigui el tunejat, es procedeix a realitzar una validació creuada amb k=10. Donat que svm() ens proporciona el MSE per a realitzar comparacions, s' ha utilitzat aquest estadístic per a fer l'anàlisi.

Per als paràmetres trobats manualment, el resultat ha estat:

```
> summary(res)

Call:
svm.default(x = enerauxtrain[, 1:24], y = enerauxtrain[, 24 + k], gam
ma = 10, cost = 2, epsilon = 0.5, cross = 10)

Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  2
      gamma:  10
    epsilon:  0.5

Number of Support Vectors:  323

10-fold cross-validation on training data:

Total Mean Squared Error: 10.21149
Squared Correlation Coefficient: 0.009533059
Mean Squared Errors:
 7.978464 12.75814 11.22317 11.23739 11.47533 11.42192 8.293724 10.21
161 9.235391 8.216413
```

Per als millors paràmetres tunejats mitjançant "grid-search", el resultat ha estat:

```
> summary(res)

Call:
svm.default(x = enerauxtrain[, 1:24], y = enerauxtrain[, 24 + k], gamm
a = 10, cost = tuned$best.parameters$cost, epsilon = tuned$best.parame
ters$epsilon, cross = 10)

Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1.05
      gamma:  10
    epsilon:  0.2

Number of Support Vectors:  374

10-fold cross-validation on training data:

Total Mean Squared Error: 9.47465
Squared Correlation Coefficient: 0.1173796
Mean Squared Errors:
 8.556841 9.054785 13.02858 9.353705 8.387982 6.238313 9.179088 10.694
4 8.739923 11.52442
```

Tot i que la diferència és poca, una MSE més baixa i una major correlació dels coeficients denota que els paràmetres trobats mitjançant "grid-search" són més òptims.

4.3.2 Model bàsic SVR

Un cop obtinguts els millors paràmetres, es procedeix a realitzar l'entrenament i així obtenir els models de regressió de vectors de suport que utilitzarem per a fer prediccions. Seguidament es realitza un test amb les 100 dades que han estat separades del conjunt d'entrenament.

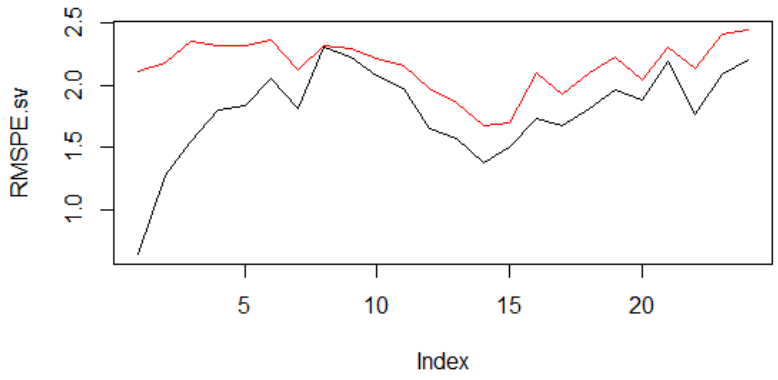


Fig. 19 Comparativa del *RMSPE resultant del models SVR i lineal amb els millors paràmetres trobats mitjançant "grid-search". Font pròpia RStudio*

S'observa que el model per a prediccions sobre el conjunt de dades de test és molt pitjor que sobre el conjunt d'entrenament. Per una banda, anteriorment s'ha vist que els models ajusten molt bé les dades d'entrenament, amb una RSE al voltant de 0,6, però per a prediccions de nous dies els resultats mostren una RMSPE al voltant de 2,2. Aquest fenomen evidència un sobre-ajust que cal solucionar. D'altre banda, veiem que el RMSPE és pitjor també que per models lineals (que són també molt dolents), quan anteriorment sobre les dades d'entrenament s'havia vist una millora important.

4.3.3 El problema del sobre-ajust

Vists els resultats de l'anàlisi anterior, el problema de modelar amb màquines de vectors de suport no recau tant en la selecció de paràmetres, sino en superar el problema de sobre ajustar les dades d'entrenament. Fins ara hem aconseguit trobar paràmetres molt bons per a fer prediccions sobre les dades d'entrenament, però no serveixen de res si no poden predir dades de hores futures.

Sobreajustar les dades significa ajustar tant al conjunt d'observacions que s'ajusta fins i tot el soroll propi d'aquestes. En aprenentatge automàtic, el sobreajust és l'efecte de sobreentrenar l'algoritme d'aprenentatge amb certes dades per als que es coneix el resultat desitjat. L'algoritme d'aprenentatge ha d'assolir un estat en el que serà capaç de predir el resultat en altres casos a partir del que s'ha après amb les dades d'entrenament, generalitzant per poder resoldre situacions diferents a les esdevingudes durant l'entrenament. No obstant això, quan un sistema s'entrena massa, l'algoritme d'aprenentatge pot quedar ajustat a unes característiques molt específiques de les dades d'entrenament que no tenen relació causal amb la funció objectiu.

Per a poder fer previsions de produccions energètiques caldrà canviar d'estratègia. És possible que calgui sacrificar certa qualitat d'ajust en les dades d'entrenament per obtenir un model més general i fer front el problema de sobre ajust, però s'haurà de tenir cura de no sota-ajustar per tal de seguir tenint un sistema eficaç.

Sabem que el paràmetre gamma propi de la funció Kernel caracteritza la complexitat de la transformació. Quant més gran és més complexa és la transformació i és possible que sigui el causant del sobre-ajust. La primera estratègia per proposada per resoldre el sobre-ajust és disminuir el valor del Kernel. D'aquesta manera sacrificarem ajust sobre les dades d'entrenament, però s'espera que el model sigui més generalitzat i ajusti millor les dades d'entrenament.

S'han realitzat diferents probes amb les mostres de test per a valors de paràmetres fixos. El

resultat ha estat el següent:

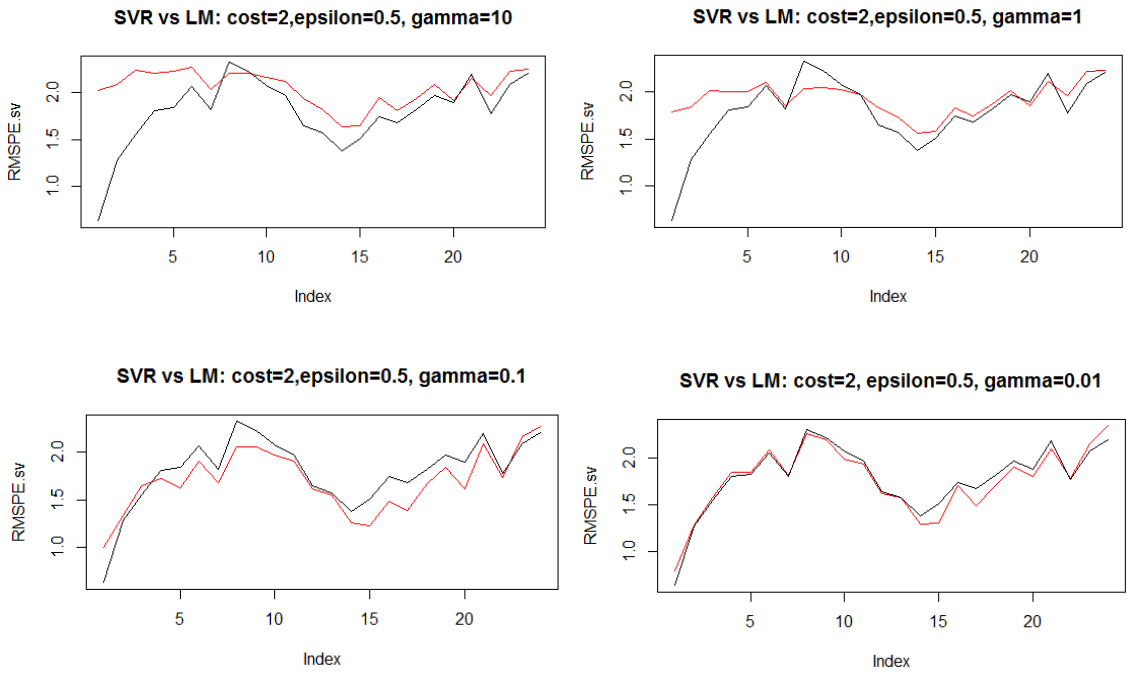


Fig. 20 Fig.17 Comparativa del RMSPE resultant del models SVR i lineal amb diferents gammas.

Tal i com s'esperava, la reducció del paràmetre gamma ha provocat la millora dels models, ja que ha aconseguit generalitzar i ajustar millor a noves dades. Per a valors de 0.1 s'ha aconseguit els millors resultats. Cal anotar que, com ha estat explicat al apartat [4.2.1.4.2], per a valors molt petits de gamma el model s'aproxima al model lineal, cosa que tampoc ens convé perquè volem obtenir un model millor que el clàssic model lineal.

A continuació, s'executa el model per al nou paràmetre gamma 0,1 amb els millors paràmetres tunejats:

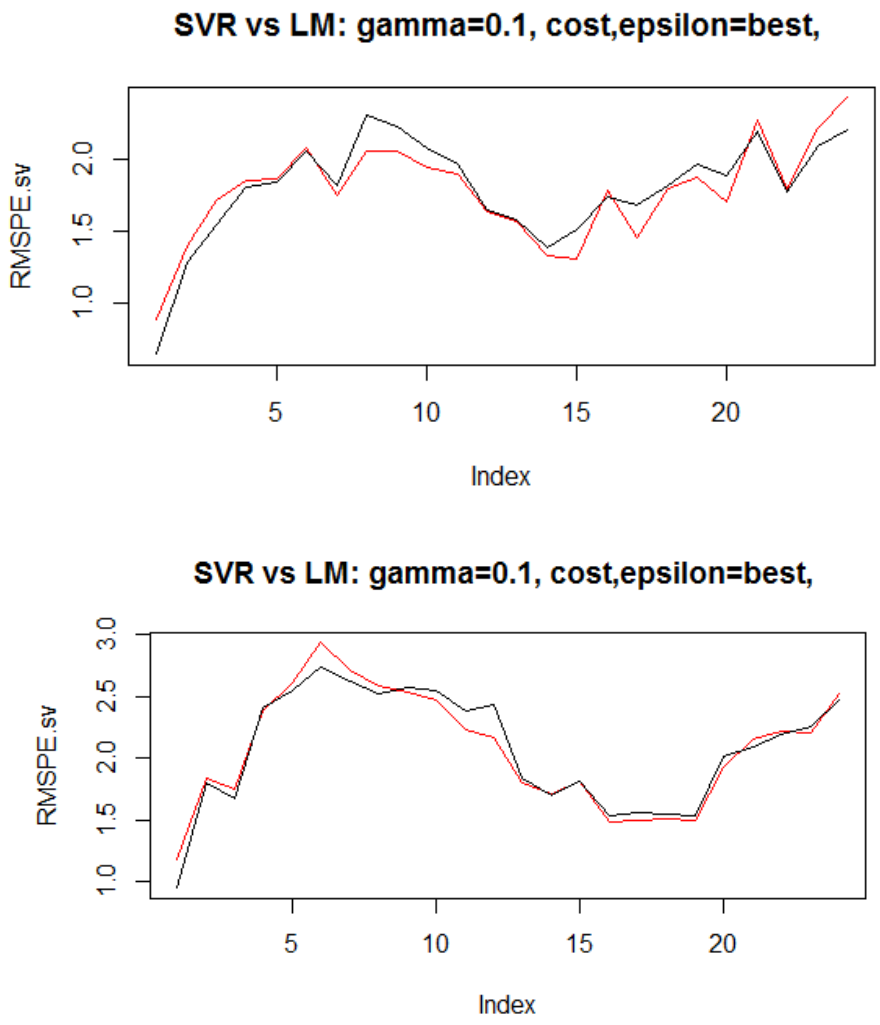


Fig. 21 Comparativa del *RMSPE resultant del models SVR i lineal, prenent diferents llavors.*

Observem que no s'aconsegueix superar l'ajust del model lineal, sinó únicament una aproximació a aquest. A més, tot i haver utilitzat paràmetres més optimitzats per a les dades d'entrenament, el resultat és molt semblant a utilitzar paràmetres fixos bons prou bons. És a dir, quan s'apropa al model lineal, tampoc es treu gaire profit de la optimització de paràmetres per grich-search.

A mode de resum, aquest sistema de predicció ajusta molt bé es dades d'entrenament, però fins i tot amb paràmetres que generalitzen els models no aconsegueix millors resultats que el model lineal per a noves observacions. Seguint aquesta estratègia no s'ha pogut aprofitar la flexibilitat de les màquines de vectors de suport, una flexibilitat que s'ha demostrat que existeix, ja que s'ha apreciat ajustant les dades d'entrenament. Al final s'ha acabat ajustant noves observacions d'una manera tant poc flexible com els models paramètrics clàssics. Es

conclou aquesta exploració comprenent que si es vol obtenir un model ajustat més acurat per a aquest tipus de series temporals s'haurà de canviar d'estratègia.

4.3.4 Model SVR ampliat

Per tal de superar el problema del sobre-ajust, s'ha vist que la modificació dels paràmetres no és suficient i per tant cal proposar alguna modificació de l'estratègia que pugui suposar uns canvis més notables.

Es proposa ampliar el sistema amb un conjunt de dades de dies anteriors, és a dir, prendre no només el dia D-1 per a l'entrenament, sinó també D-2. S'ha descartat la idea de prendre dades més llunyanes (D-3, D-4, etc.) ja que al tractar-se de vent, es coneix que a més de dos dies vista la dependència és molt baixa.

Seguint el mateix procés que en l'estratègia original es torna a realitzar l'ajust. A la Fig. 22 i Fig. 22 s'han graficat diferents trams consecutius de la sèrie temporal per tal d'observar el que realment aconseguim. L'eix horitzontal representa les hores, de les quals les 48 primeres són corresponents a les dades dels dies D-2 (0-24h) i D-1 (24-48h). L'eix vertical és la potència generada aquella hora. La Fig. 22 representa l'entrenament i la Fig.23 representa el test.

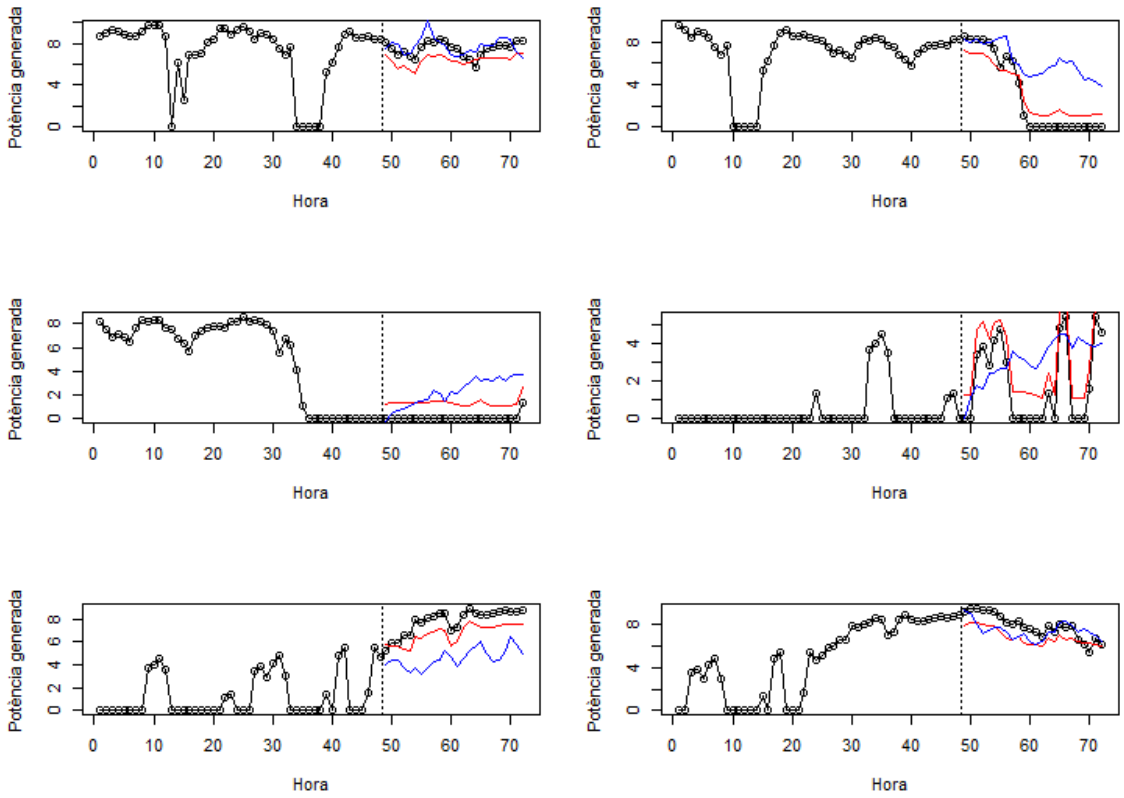


Fig. 22 *Sèrie temporal de dies successius. Les produccions horàries en negre, la predicció del test que proporcionen svm en vermell i models lineals en blau.*

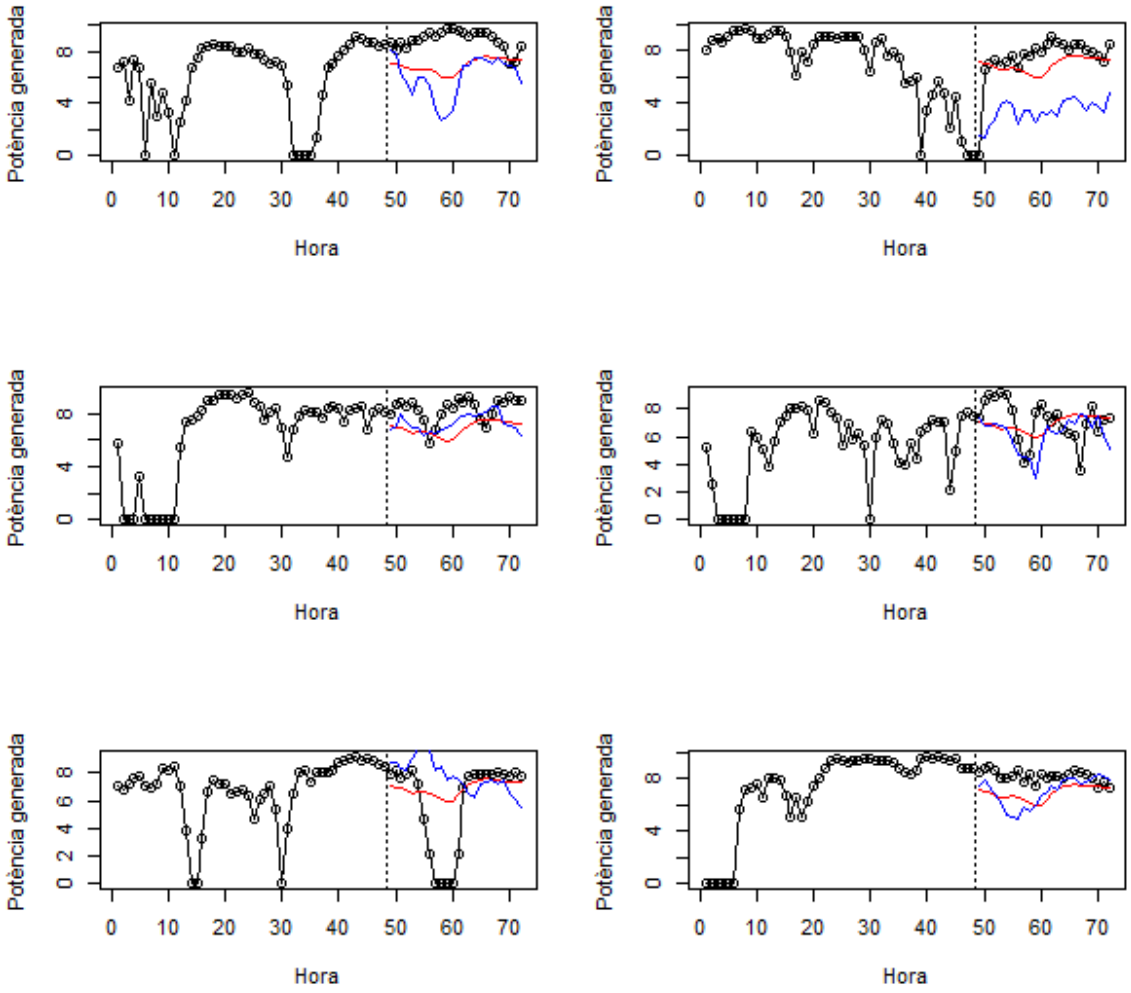


Fig. 23 *Sèrie temporal de dies successius. Les produccions horàries en negre, la predicció del test que proporcionen svm en vermell i models lineals en blau.*

Observant el comportament de l'ajust per a dades d'entrenament, s'aprecia com encara els resultats del SVR són molt més satisfactoris que amb models lineals. La seva flexibilitat permet que s'adapti molt bé a canvis bruscos i identifiqui molt bé els patrons de la sèrie. En canvi, els models lineals són molt més rígids i només ofereixen precisió quan les dades ajustades són molt properes a observacions anteriors.

No obstant, els resultats continuen no sent satisfactoris. Es veu com tant els models lineals com svm són massa rígids per predir noves observacions. Els canvis bruscos de vent continuen sense poder ser predits i la s'ha perdut la flexibilitat que caracteritza l'entrenament. L'ampliació dels models amb dades anteriors no ha estat suficient per a millorar els resultats.

4.3.5 Model SVR entrenat amb suavització

Donat que s'ha vist que la major font de problemes per a l'ajust és la irregularitat de les dades que provoca canvis bruscos difícils de predir, es proposa un mètode de suavització. Els mètodes de suavització eliminen les fluctuacions aleatòries de la sèrie de temps, proporcionant dades menys distorsionades del comportament.

El sistema utilitzat serà el mateix que l'anterior, únicament amb la modificació del conjunt de dades d'entrenament. S'espera que realitzant suavitzacions, el sistema sigui perdi rigidesa i els models siguin més flexibles alhora de predir noves produccions.

Observant el resultat de l'entrenament per a alguns dies al atzar:

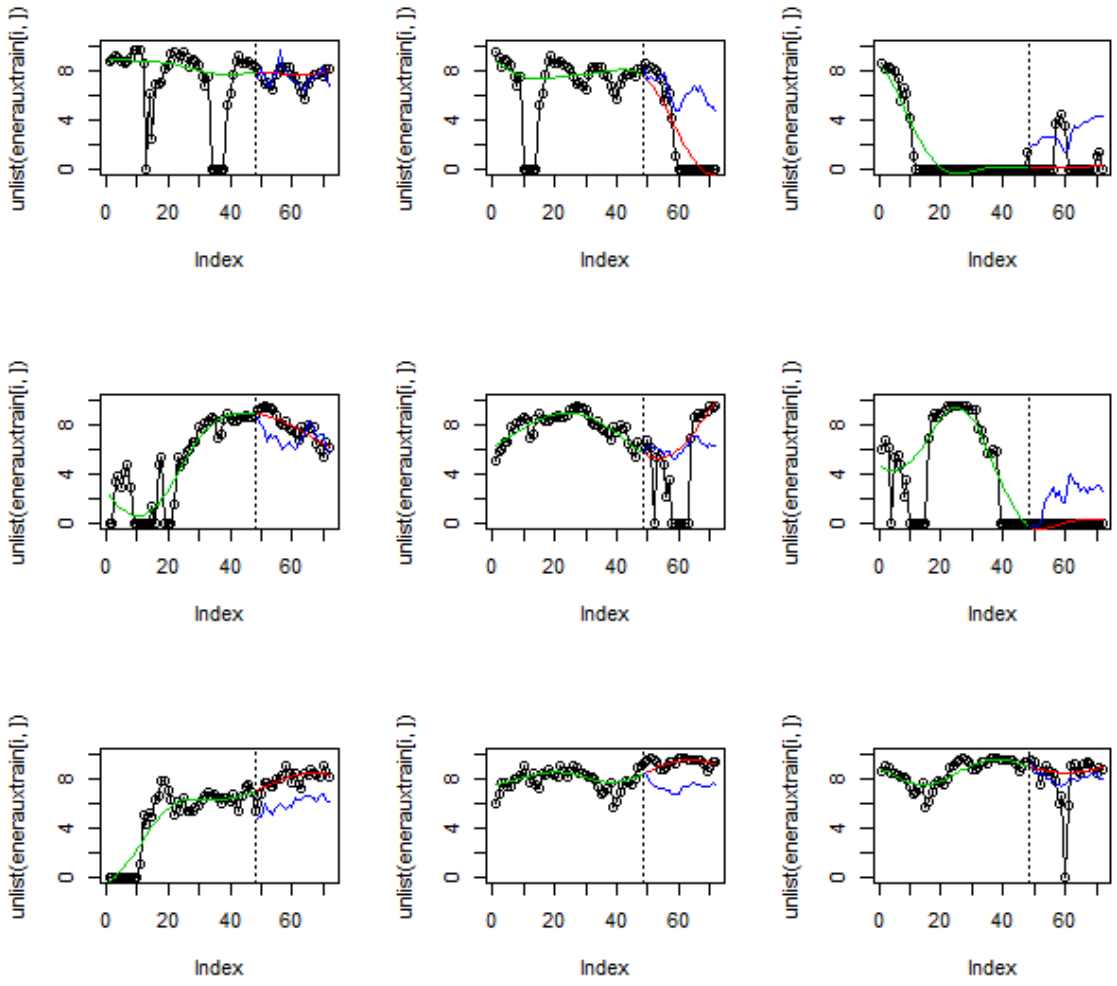


Fig. 24 Sèrie temporal de dies successius. Les produccions horàries en negre, la producció suavitzada en verd, la predicció del test svm en vermell i models lineals en blau.

Com es pot veure en la Fig.25, l'ajust de les dades d'entrenament del model SVR és tant bo com la pròpia suavització (el model és la continuació de la mateixa suavització) i l'únic error respecte les produccions és el propi caràcter irregular d'aquestes. L'ajust és immillorable. El model lineal en canvi es desvia de forma semblant al sistema anterior sense suavització, ja que per cada hora realitza un ajust diferent independent la l'ajust de la hora anterior.

Observant el resultat del test per a alguns dies al atzar:

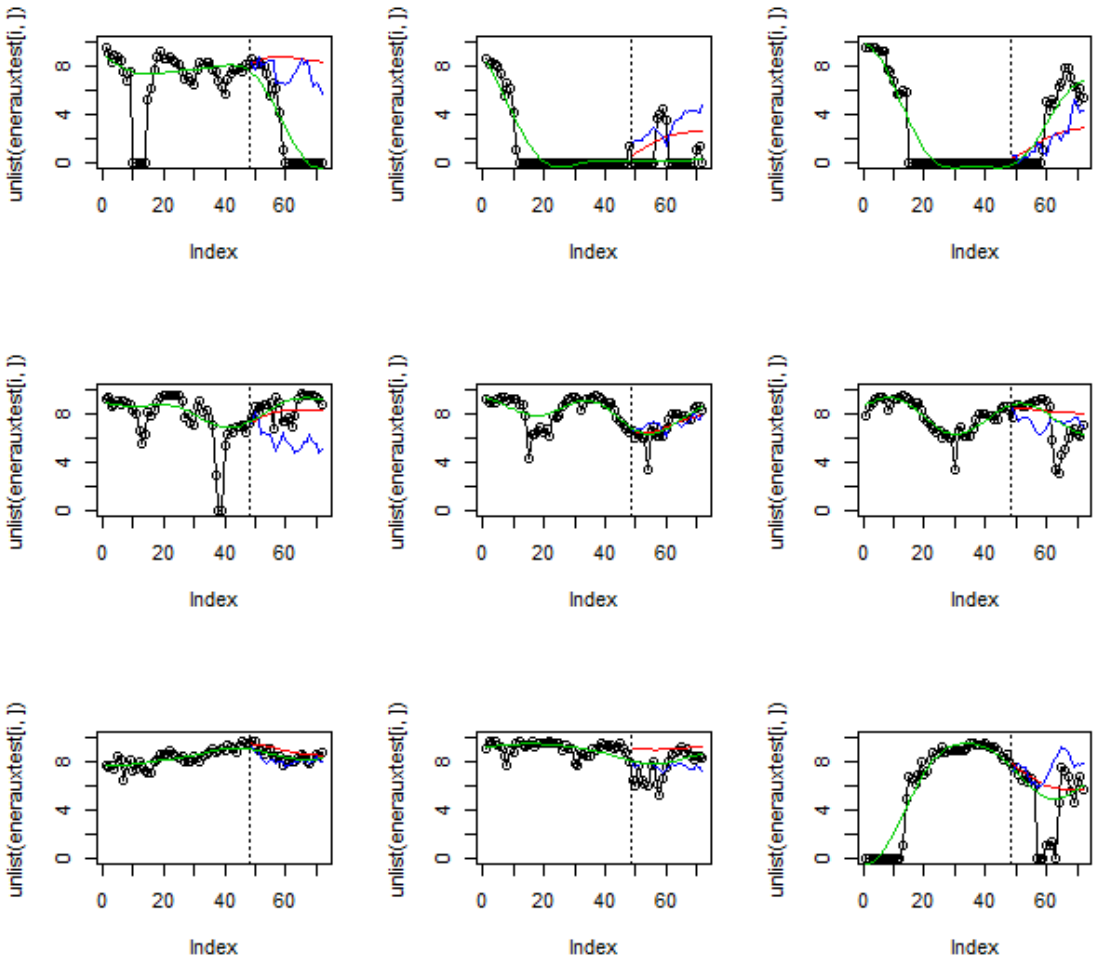


Fig. 25 Sèrie temporal de dies successius. Les produccions horàries en negre, la producció suavitzada en verd, la predicció del test svm en vermell i models lineals en blau.

Es pot observar en els dos gràfics de l'esquerra de la segona i tercera fila de la Fig.26 que les màquines de vectors de suport prediuen força bé quan la producció és més o menys estable. Per contra, quan hi ha canvis bruscos de produccions, el sistema és incapaç de predir aquest resultats ja que la fluctuació és massa irregular i els canvis són massa sobtats.

És molt destacable que els resultats del tests mostren que els patrons seguits dels models

són molt menys rígids que anteriorment sense la suavització. Aquest resultats són per aquest motiu molt satisfactoris, ja que justament l'objectiu de la suavització era aconseguir flexibilitat a les proves de test.

Els errors resultants dels models SVR comparats amb els models lineals són:

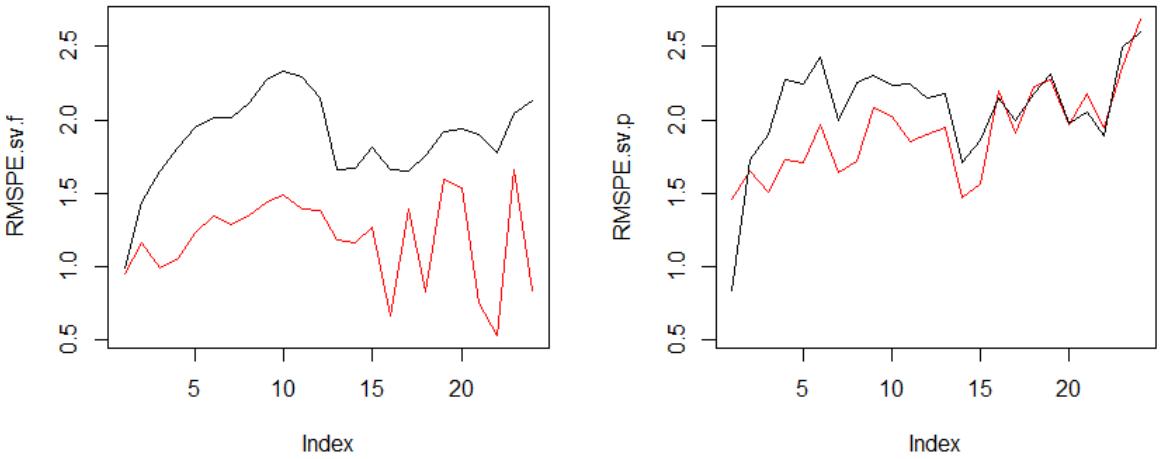


Fig. 26 Comparativa del RMSPE resultant del models SVR i lineal amb suavització. Entrenament a la dreta i test a l'esquerra.

La conclusió és que s'ha perdut cert ajust sobre les dades d'entrenament, que anteriorment donava una RMSPE mitjana de les 24 hores al voltant de 0.6 i actualment al voltant de 1. A canvi s'ha guanyat certa flexibilitat que s'ha vist reflectida en una disminució de l'error de predicció al test, que és el que realment interessa.

Es pot afirmar que el sistema suavitzat és millor que el sistema sense suavització. Gràcies a aquest a tècnica s'ha assolit l'objectiu de trobar un sistema basat en màquines de vectors de suport millor que els models lineals.

4.3.6 Altres estratègies explorades

De forma resumida, en aquest apartat s'expliquen altres estratègies explorades durant el desenvolupament de la implementació, pero que no han estat satisfactories per certs motius.

En primer lloc, una tècnica explorada ha estat el "train-test-validation". La idea consisteix en separar el conjunt de dades, no en dues parts (entrenament i test) com fins ara, sinó en tres conjunts: entrenament, validació i test. Es tracta de tenir un conjunt anomenat validació que permeti ens estudiar l'error comès quan es realitzen prediccions sobre noves observacions, és a dir, que permeti extreure informmació del que fins ara era test.

L'ús d'aquesta mètode consisteix en probar els diferents paràmetres a l'entrenament per tal de trobar un punt de convergencia óptim entre l'ajust a l'entrenament i la predicció a la validació i així poder fer posteriorment el test amb més possibilitats d'obtenir precisió.

No obstant, el resultat tampoc ha estat l'esperat. En aquest cas particular s'ha vist que l'entrenament i la validació no convergeixen mai perquè els residus sempre són a escales diferents com es pot apreciar a la figura següent. A la Fig. 28 es mostra un exemple en que el RMSPE per a l'entrenament sempre pren valors baixos per a paràmetres acceptables, en aquest cas concret està al voltant de 0.6, i altres paràmetres es dispara massa alt. El RMSPE per a la validació está al voltant del 2.5 per paràmetres acceptables, i del contrari també es dispara massa alt. És a dir, no hi ha un valor on l'entrenament i la validació convergeixin. En consecuencia no cal ni que es fagi test. Aquest mètode no ha resultat eficaç.

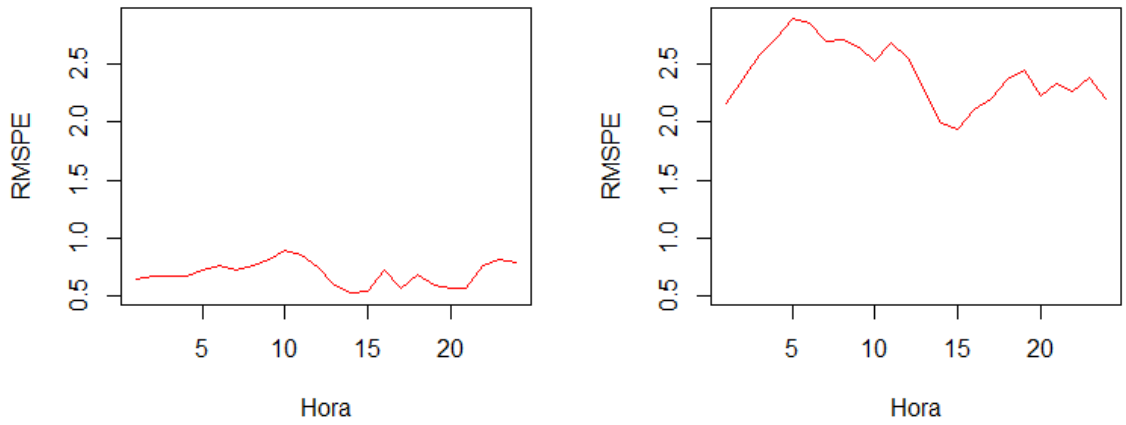


Fig. 27 *Entrenament a la dreta i validació a l'esquerra amb paràmetres òptims de SVR.*

En segon lloc, també s'ha plantejat realitzar una classificació de les dades per tal d'afegir informació al nostre sistema. La idea consisteix en identificar diferents classes en el conjunt de dades, per tal de posteriorment segmentar les dades en diferents grups on les dades siguin similars i aplicar el sistema de SVR per a cada grup. D'aquesta manera l'algorisme de màquines de vectors de suport té més facilitats per identificar els patrons.

Donat que no es té coneixement a priori d'aquestes classes, queda descartada directament la opció d'emprar l'aplicació de les propies màquines de vectors de suport per a la classificació. No obstant, existeixen molts mètodes de classificació no supervisada (que no requereixen d'informació a priori) per a realitzar la classificació. El mètode que s'ha explorat ha estat l'anàlisi clúster per agrupament jeràrquic, que està basat en la idea principal que els objectes més propers estan més relacionats que els que estan allunyats i que permet

desglosar el conjunt de dades per branques i identificar diferents els grups.

Malauradament, al aplicar la tècnica i identificar els diferents grups, el conjunt de dades queda molt reduït, i els petits subconjunts són insuficients per a entrenar el model de màquines de vectors de suport.

Com a conclusió d'aquest apartat, hi ha altres mètodes molt interessants que podrien superar el problema del sobreajust del sistema obtingut, però la limitació de les poques dades que disposem resulta un impediment alhora de trobar un sistema solució, ja que moltes de les eïnes existents requereixen quantitats ingents de dades per a mostrar resultats satisfactoris.

4.4 Resultats del sistema definitiu

Finalment, després de desenvolupar diverses millores, s'ha arribat al sistema de previsions horàries d'energia eòlica basat en màquines de vectors de suport per a regressió. En aquest apartat es resumeix els trets característics d'aquest sistema solució final i s'exposen els seus resultats mitjançant una validació creuda.

El sistema solució consisteix en un conjunt de models predictius de les produccions energètiques de les 24 hores del dia D, alimentat per la suavització de les dades d'entrenament de les produccions energètiques dels dies D-1 i D-2.

Per tal de poder avaluar la precisió del sistema, s'exposen els resultats de l'ajust dels entrenaments i de les prediccions del test i a la vegada es mostren comparativament amb els models lineals. Mitjançant l'estimador RMSPE es quantifica aquest error i comparen els resultats dels models per a les 24 hores del dia D. Per garantir que l'avaluació és fiable, s'aplica la tecnica de validació creuada "k-fold cross validation", segmentant les dades en cinc parts (k=5) i utilitzant sistemàticament una d'elles com a test i les altres 4 com a entrenament, realitzant així un total de 5 probes de validació.

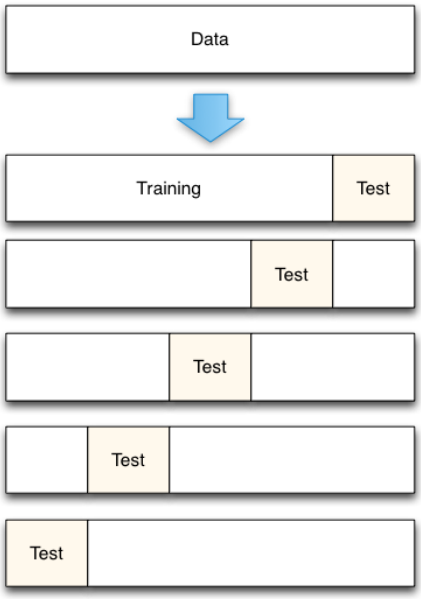
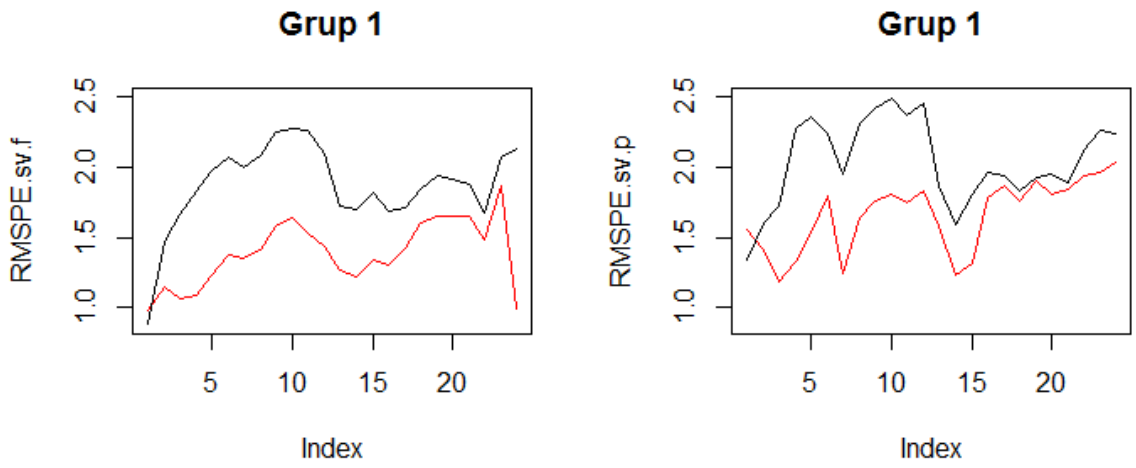
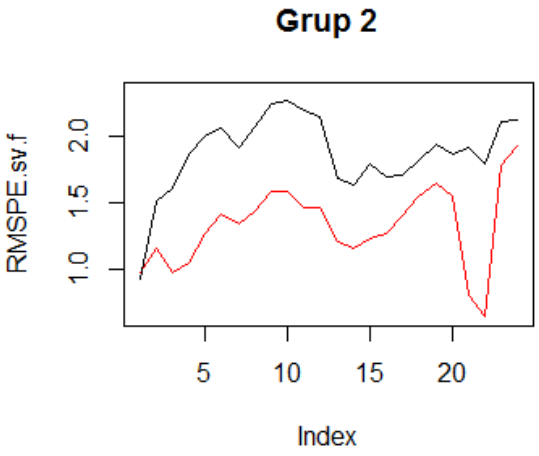


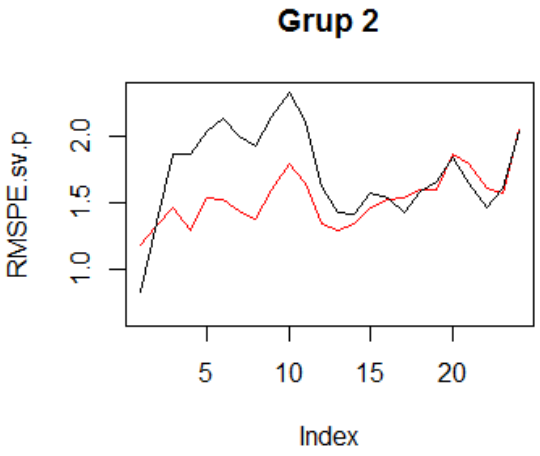
Fig. 28 *K-fold cross validation amb k=5. Font: Kaggle*

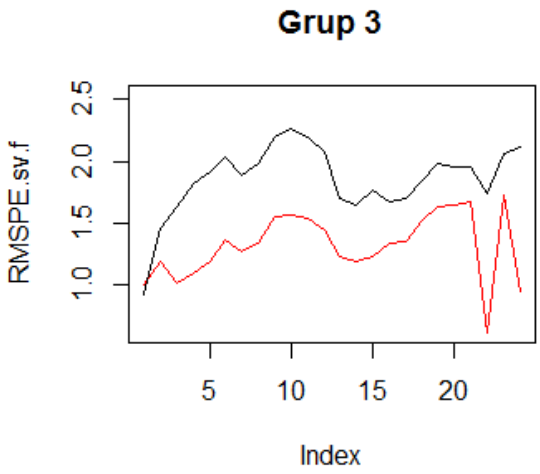
Amb aquest anàlisi de precisió es pretèn donar una informació honesta de com acurat és el sistema predictiu.

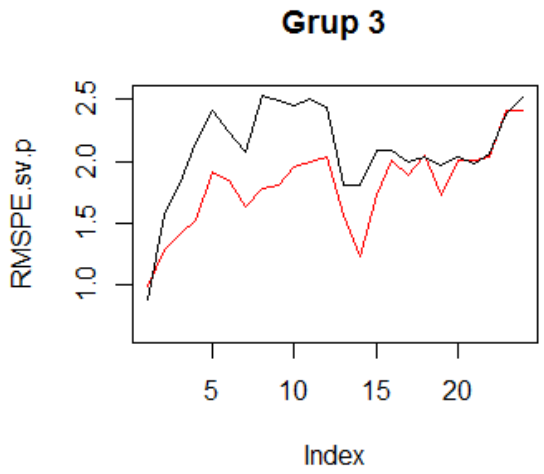
El resultat per aquestes 5 probes són:

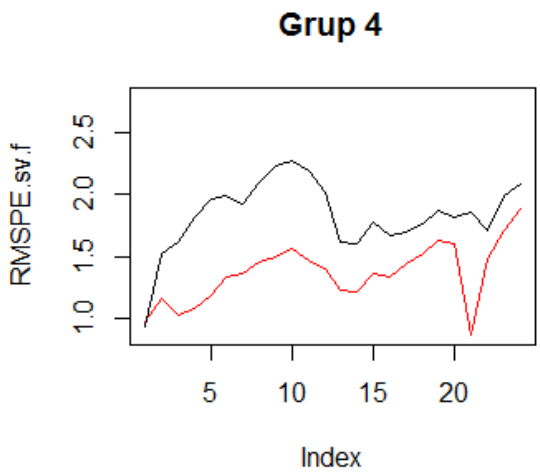


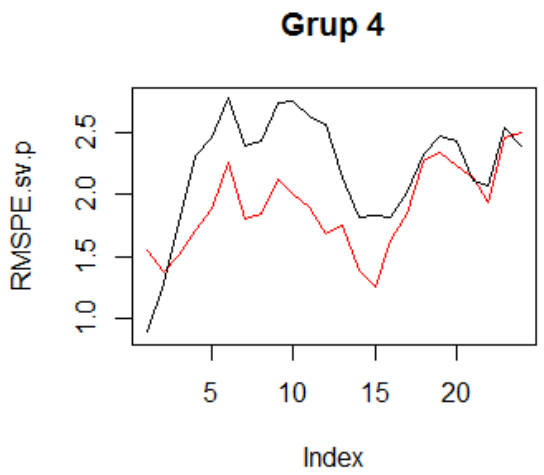

Grup 2
RMSPE.sv.f
1.0
1.5
2.0
5
10
15
20
Index


Grup 2
RMSPE.sv.p
1.0
1.5
2.0
5
10
15
20
Index


Grup 3
RMSPE.sv.f
1.0
1.5
2.0
2.5
5
10
15
20
Index


Grup 3
RMSPE.sv.p
1.0
1.5
2.0
2.5
5
10
15
20
Index


Grup 4
RMSPE.sv.f
1.0
1.5
2.0
2.5
5
10
15
20
Index


Grup 4
RMSPE.sv.p
1.0
1.5
2.0
2.5
5
10
15
20
Index

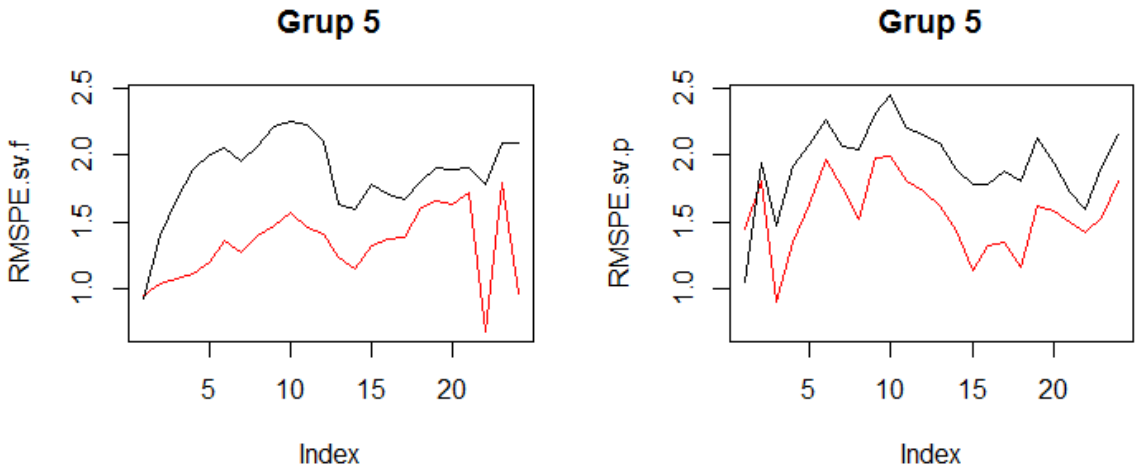


Fig. 29 Resultats de les cinc probes de validació. RMSPE del sistema de previsions horàries d'energia eòlica en vermell, RMSPE dels models lineals en negre. A la esquerra ajust d'entrenament i a la dreta prediccions de test.

És important recordar que el que realment interessa per realitzar prediccions són les probes de test, ja que per a la construcció dels models no s'utilitza la seva informació, i actuen com a noves dades futures amb les que comparem la predicció dels models.

Com es pot observar en les 5 probes, els resultats de les prediccions proporcionades per el sistema basat en màquines de vectors de suport són força més precisos que el model lineal. Durant les 15 primeres hores el sistema és força més acurat i a les últimes hores les prediccions són molt semblats a les del model lineal. És lògic que les últimes hores les prediccions siguin menys precises ja que és més difícil realitzar prediccions d'horitzons llunyans. Únicament la primera hora del dia és més ben predita pels models lineals.

Per últim, comentar que tot i que els resultat són satisfactoris per haver arribat a un sistema millor que els models lineals, sent ambiciosos i crítics es pot que afirmar que no s'ha arribat encara a un sistema predictiu acurat. De fet, s'està considerant un error RMSPE superior a la unitat que, al estar parlant d'un estimador de l'arrel quadrada dels errors relatius, suposa que sigui una mesura força alta. A més s'ha de tenir en compte que les dades sobre les que s'ha treballat són una transformació logarítmica, cosa que augmenta exponencialment l'error real.

5. Conclusions

A mode de resum, per a la elaboració d'aquest treball s'ha realitzat una documentació del funcionament i la importància de la producció d'energia eòlica a España, per tal de poder justificar la necessitat de les previsions energètiques. Posteriorment s'ha dut a terme una revisió teòrica dels principis de les màquines de vectors de suport tant per a regressió com per a classificació. Una vegada familiaritzats amb el problema i l'eina de treball, s'ha implementat SVR per a la creació de sistemes que realitzen models predictius amb un horitzó de 24 hores, en funció de dades de fins a 48 hores anteriors. Així, s'ha pogut explorar el comportament de les màquines de vectors de suport i presenciar els seus avantatges i inconvenients.

A continuació es presenten les conclusions principals, les aportacions originals del treball i les futures línies d'investigació que es proposen.

En primer lloc, els mètodes SVM són avantatjosos perquè, gràcies a les transformacions del Kernel, permeten molta flexibilitat alhora de ajustar models de regressió, ja que no necessiten que les observacions descriguin formes conegudes per ajustar les dades com pot passar en els models paramètrics clàssics. Aquest fet s'ha apreciat satisfactòriament en el seu entrenament, on s'ha observat que l'ajust és molt bo i els errors són molt baixos.

No obstant, a les màquines de vectors de suport, la gran virtud de flexibilitat pot ser a la vegada el seu major inconvenient ja que és la causa del sobre-ajust. Aquest fenomen ha quedat evidenciat veient que els valors predits de les dades d'entrenament s'adeqüen a les dades amb relativa facilitat, mentre que els valors predits de les dades de test han mostrat resultats molt poc precisos en tots els casos. S'han provat diferents mètodes per tal de combatre aquest fenomen però l'eina de regressió no ha estat suficient per sí sola per a consolidar un model predictiu de produccions energètiques.

Les màquines de vectors de suport sobre-ajusten amb facilitat i per això és perillós utilitzar-les sense conèixer bé el seu funcionament. Sovint s'empren com a caixa negra i es prenen els seus resultats d'entrenament amb gran satisfacció, sense consciència de que ajustar molt bé l'entrenament no serveix de res si el test no és apropiat. Unes observacions sobre-ajustades poden servir com a eina de interpolació, però no com a sistema de previsions. Amb la finalització d'aquest treball s'aporta a la investigació que en la aplicació de màquines de vectors de suport per a regressió no sempre es tant important l'ajust òptim dels paràmetres per a l'entrenament, fet en el que es centren la majoria de documents referits a la implementació de màquines de vectors de suport , sinó sovint és més important la capacitat de trobar l'error mínim irreductible en un model ni sobre-ajustat, ni sota-ajustat.

Respecte a les dificultats que s'han trobat per al desenvolupament del sistema, han estat: Per una banda, el propi caràcter irregular de les dades de produccions energètiques, no només pel sentit meteorològic que implica treballar amb dades derivades del vent, sinó també per les condicions afegides pels aerogeneradors, que afegeix davallades imprevistes de la producció. D'altre banda, la poca quantitat de dades de les que s'ha disposat per a ajustar models no paramètrics. Es coneix que per a obtenir resultats satisfactoris al món de l'aprenentatge automàtic es necessiten moltes dades i són realment útils quan es disposen de quantitats ingents.

Conclòs aquest treball encara no s'ha pogut establir un sistema de previsions precís, però s'ha explorat la metodologia d'ús de SVM i els seus resultats. Que no hagi funcionat com idealment podíem esperar no significa que s'hagi de descartar aquesta tècnica, sinó que potser s'ha de complementar amb altres eines. Per aquest motiu, s'obren noves futures línies d'investigació.

En primer lloc, es podria proposar utilitzar la mateixa tècnica de regressió de màquines de vectors de suport però afegint les mateixes dades predites al conjunt d'entrenament. Així, a la segona hora del dia D s'inclouria al conjunt d'entrenament la predicció de la primera hora, a la tercera hora, la predicció de la primera i de la segona hora, i així successivament. Aquesta tècnica podria resultar útil per a un sistema en què les prediccions per a la primera hora del dia D fossin molt bones, ja que s'afegiria informació rellevant.

En segon lloc, es podria proposar utilitzar diferents mètodes estadístics predictius i acoblar els seus resultats de manera que els resultats mostrats fossin els millors resultats ofereixin en les prediccions. Aquest sistema podria resultar interessant ja que es podria utilitzar mètodes no paramètrics com SVM i mètodes paramètrics per a series temporals com els mètodes ARIMA, i a la vegada es podrien combinar models estadístics i models físics. Tot plegat, probablement donaria molt bones prediccions, però segur el sistema resultaria molt complex de construir.

Per últim, una molt bona opció seria combinar altres tipus de dades. En els sistemes utilitzats per aquest treball només s'ha utilitzat dades numèriques de les produccions energètiques del parc eòlic, però el fet d'estar estrictament lligades a la condicions meteorològiques del vent resulta evident que utilitzar registres del vent aportaria informació determinant. Tenint les dades del vent es podria identificar si els zeros de les observacions són propis d'aturades del aerogenerador o de nivells baixos de vent i d'aquesta manera afegir informació al sistema alhora de segmentar les dades. Si no s'ha incorporat aquest estudi ha estat perquè no es disposa dels registres del vent corresponents al conjunt de dades analitzat, ja que no es coneixen les dates de la font i només es coneix el seu ordre cronològic.

6. Impacte del Projecte

En aquest apartat del projecte s'exposen els costos i beneficis que ha suposat la realització del projecte, així com l'impacte mediambiental que ha tingut i tindria si fos utilitzat a gran escala.

6.1 Impacte Econòmic

6.1.1 Costos del projecte

Respecte als costos que ha suposat la realització, poden desglossar-se en costos de infraestructura, consum energètic i recursos humans.

Els costos principals han estat de recursos humans, associats a hores de treball. S'estima que han estat invertides unes 400 hores de treball per a la realització del projecte, que inclouen tant les hores de familiarització amb el problema i aprenentatge del mètode, com la pròpia implementació de les eines utilitzades i el redactat del informe. A part, també hi ha els costos associats a l'ajuda del director del treball, referides a les hores de les reunions amb l'estudiant.

Els costos propis de infraestructura inclouen els emplaçaments on s'ha dut a terme el projecte, que ha estat bàsicament la propietat de l'estudiant, la universitat on estudia i, puntualment altres biblioteques. Així mateix, els consums energètics de llums d'aquests establiments també s'inclouen. Físicament el projecte s'ha desenvolupat per mitjà d'ordinador, que també ha implicat un consum energètic. Tots aquests costos d'infraestructura són negligibles ja que no han estat un cost directe per l'estudiant i seria un cost igualment produït si no s'hagués efectuat el treball. Potser l'únic cost que es podria contemplar és el del consum energètic d'hores treballades a la propietat de l'estudiant.

El conjunt de dades amb el que s'ha treballat ha estat facilitat per Gas Natural i no ha suposat cap cost per a l'estudiant.

6.1.2 Beneficis del projecte

Donat que el projecte és una iniciativa independent al mercat elèctric i que engega des de zero un sistema de previsions d'energia eòlica, no podem parlar de que hagi generat cap benefici econòmic fins al moment. Tampoc era aquesta la intenció del projecte. No obstant, podem dir que si es continues treballant en el projecte, podria aportar notables beneficis econòmics per al productor.

Com s'ha explicat en en capítol 2, el valor de les previsions de produccions energètiques per al mercat elèctric és molt alt. Per una banda, ajuda a minimitzar les possibles penalitzacions econòmiques que s'apliquen als agents a causa de les desviacions en els seus compromisos de generació adquirits en el mercat d'energia elèctrica. D'altra banda, afecta en la reducció de costos d'operació en el sistema originada per la reducció de reserva elèctrica necessària.

6.2 Impacte Mediambiental

Els sistemes elèctrics presenten com a problema característic la impossibilitat actual de poder emmagatzemar energia a gran escala, fet pel qual requereix que l'oferta sigui igual a la demanda en cada instant de temps, el que suposa necessàriament una coordinació de la producció d'energia elèctrica, així com la coordinació entre les decisions d'inversió en generació i en transport d'energia elèctrica.

D'altre banda, un sistema elèctric fiable ha d'assegurar que la demanda d'energia elèctrica per part dels usuaris, estigui suficientment coberta per la generació elèctrica disponible a cada moment.

D'aquesta manera es fa evident que disposar de previsions de produccions elèctriques tindrà un impacte important en la gestió elèctrica, és a dir, en els sistemes de generació, transport i distribució d'energia elèctrica, ja que permet la determinació de la reserva de la generació i la bona qualitat i seguretat en el subministrament elèctric.

A tall de conclusió, crear sistemes de predicció de produccions energètiques provoca una millora de la gestió de l'electricitat provinent d'aerogeneradors, fet que tindrà repercussions molt positives en termes mediambientals, ja que permetrà treure el màxim profit de les produccions energètiques renovables i per tant estimular aquest sector.

Agraïments

Vull aprofitar aquestes línies per agrair al director d'aquest treball, Josep Antón Sánchez, per haver-me proposat la realització d'aquest treball i obrir-me les portes al món del machine learning, i així mateix haver mostrat el seu continu suport, orientació i consells en el seu desenvolupament.

Bibliografia

Referències bibliogràfiques

[1] Abe, Dr S. ; Professor Sameer Singh, PhD (Hrsg.): *Support Vector Machines for pattern Classication*. Springer, 2005

[2] V. Vapnik. The Nature of Statistical Learning Theory. Springer, N.Y., 1995

[3] V. Vapnik, S. Golowich, and A. Smola.Support vector method for function approximation, regression estimation, and signal processin.

[4] Cristianini, Nello ; Shawe-Taylor, John: An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.

[5] A. Smola and B. Scholkopf: *A Tutorial on Support Vector Regression*. October 1998

[6] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. *A Practical Guide to Support Vector Classication*. Department of Computer Science, National Taiwan University.

[7] David Meyer. Package *'e1071', August 2015*. [https://cran.r-project.org/web/packages/e1071/e1071.pdf]

Bibliografia complementària

[8] Alex J Smola and Bernhard Scholkopf. *Math Tutorial on Support Vector Regression.* http://www.svms.org/regression/SmSc98.pdf

[9] AJ Smola and Schlkopf. *From regularization operators to support vector kernels In Advances in Neural information processings systems.* San Mateo 1998.

[10] Lee, Y.; Lin, Y. & Wahba, G. *Multicategory Support Vector Machines*. 2001.

[11] RPubs*. Support Vector Regression.* [http://rpubs.com/joser/svm]

[12] Tripod*. Support Vector Regression.* [http://kernelsvm.tripod.com]

[13] L'Université de Lyon. *Tanagra i R Support Vector RegressionTutorial.* [http://eric.univlyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Support_Vector_Regression.pdf*]*

[14] Alexandre Kowalczyk. *Support Vector Regression Tutorial.* [http://www.svm-tutorial.com/2014/10/support-vector-regression-r/]

[15] Andrew Ng, Associate Professor, Stanford University. *Machine Learning Course*. [https://es.coursera.org/learn/machine-learning] 2010

[16] Andrew Ng, Associate Professor, Stanford University. *Suppot Vector Machine Lecture Notes.2010*

[17] Roger D. Peng, PhD, Associate Professor. R Programming. [https://es.coursera.org/learn/r-programming]

[18] R. Berwick, Village Idiot, An Idiot's guide to Support vector machines (SVMs). [http://www.svms.org/tutorials/Berwick2003.pdf]

[19] Alba Castro, José Luis. *Máquinas de Vectores Soporte (SVM)*. 2015.

[20] Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001)

[21] Cristianini, Nello; and Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[22] OMIE. Mercado Eléctrico Español. [http://www.omie.es/inicio]

[23] MIBEL. Mercado Ibérico de la Electricidad [www.mibel.com/]